OECD publishing

EXPLORING POSSIBLE AI TRAJECTORIES THROUGH 2030

OECD ARTIFICIAL INTELLIGENCE PAPERS

February 2026 **No. 55**

GPAI

OECD
BETTER POLICIES FOR BETTER LIVES

OECD Artificial Intelligence Papers

Exploring possible AI trajectories through 2030

Hamish Hobbs, Dexter Docherty, Luis Aranda, Kasumi Sugimoto, Karine Perset, Rafał Kierzenkowski

Note to Delegations:

This document is also available on O.N.E Members & Partners under the reference code:

DSTI/DPC/GPAI(2025)/13/FINAL

Cover image: © Kjpargeter/Shutterstock.com

Abstract

Artificial intelligence (AI) has advanced rapidly in recent years, with systems becoming increasingly capable. This paper presents expert- and evidence-informed scenarios for how AI could progress by 2030. It considers recent trends in AI and key uncertainties for AI progress through 2030. Current evidence suggests that four different broad scenario classes are all plausible through to 2030: progress stalling, progress slowing, progress continuing, and progress accelerating. This suggests that AI progress by 2030 has a plausible range that includes both a plateau at approximately today's level of capabilities and rapid improvement that leads to AI systems which broadly surpass human capabilities. This paper decomposes plausible AI capability progress in each scenario in line with the OECD's beta AI capability indicators, exploring plausible capability trajectories for AI system's abilities in language; social interaction; problem solving; creativity; metacognition and critical thinking; knowledge, learning and memory; vision; physical manipulation; and robotic intelligence.

Acknowledgements

This paper was drafted by Hamish Hobbs from the OECD Strategic Foresight Unit, in close collaboration with Dexter Docherty from the Strategic Foresight Unit and Kasumi Sugimoto, Luis Aranda and Karine Perset from the OECD Division on AI and Emerging Digital Technologies. Strategic direction and input were provided by Rafał Kierzenkowski, Senior Counsellor for Strategic Foresight and Jerry Sheehan and Audrey Plonk, respectively Director and Deputy Director of the OECD Directorate for Science, Technology and Innovation (STI).

The team gratefully acknowledges the input of Stuart Elliot, Sam Mitchell and Zina Efchary regarding the integration of the OECD beta AI Capability Indicators. The team also thanks Niamh Higgins-Lavery from the Strategic Foresight Unit for operational support and Shellie Laffont, Christy Dentler and Andreia Furtado from STI Communications and Romy de Courtay (external editor) for editorial support.

The paper benefitted significantly from the oral and written contributions of GPAI delegates as well as experts from the OECD.AI network of experts. The authors would like to extend their sincere gratitude to the Delegations of Brazil, Greece, India, Israel, Spain, Saudi Arabia, Slovenia, Türkiye, and the United Kingdom for their invaluable insights. The authors thank the members of the OECD Expert Group on AI Futures for their insightful comments.

This report benefited greatly from discussions and input from the writing team of the International AI Safety Report, including Carina Prunkl, Stephen Clare, Maksym Andriushchenko, Patrick King and Hannah Merchant.

Finally, the authors gratefully recognise the substantial contributions from external experts, including Álvaro Soto (Pontificia Universidad Católica de Chile), Friedrik Heintz (Linköping University), Gopal Ramchurn (University of Southampton), Hiroshi Ishiguro (Osaka University), Nick Jennings (Loughborough University) Jonas Sandbrink (AI Security Institute), Stuart Russell (University of California, Berkeley), Susan Leavy (University College Dublin), and Yoshua Bengio (University of Montreal).

Table of contents

Executive summary

Artificial Intelligence (AI) has advanced rapidly in recent years, with systems becoming increasingly capable. Understanding how AI might evolve by 2030 will help governments craft policies that capture the benefits of AI progress and prepare for its potential impacts.

The OECD has developed expert and evidence informed scenarios for how AI could progress by 2030, building on the OECD's beta AI Capability Indicators (OECD, 2025[1]). While the state of AI capabilities in 2030 cannot be predicted with certainty, governments would benefit from understanding a range of plausible development trajectories.

Policymakers could consider four different broad scenario classes that are all plausible by 2030:

- **Progress Stalls:** A scenario in which progress in the most advanced AI systems largely halts and capabilities remain largely unchanged. Rapid gains observed over recent years stop and AI progress plateaus. Diffusion and application development continue for existing capabilities. In 2030, AI systems can quickly undertake a range of tasks that would take humans hours to perform, but issues of robustness and hallucinations impact reliability. AI systems typically rely upon substantial support from humans to complete tasks, such as detailed prompting, review and provision of context.
- **Progress Slows:** A scenario in which incremental gains in the most advanced AI systems deliver continued but slower progress. In 2030, AI systems have a deep knowledge base, excel at standard forms of structured reasoning, and can act as useful assistants for tasks that require them to use a computer, navigate the web or undertake limited interaction with people or services on behalf of the user. AI systems can quickly undertake well-scoped tasks that would take humans hours or days to perform. AI systems typically rely on humans to provide them with clearly scoped tasks, review important decisions or actions, and provide detailed guidance and context.
- **Progress Continues:** A scenario in which continued rapid progress occurs. In 2030, AI systems can perform many professional tasks in digital environments that might take humans a month to complete. Deficits in AI system's continual learning and generalisation to complex real-world environments and situations persist. AI systems typically rely on humans to provide high level directions and bounds for their behaviour, but can often operate with high autonomy within these bounds towards a given objective, including autonomously interacting with a range of stakeholders.
- **Progress Accelerates:** A scenario in which dramatic progress leads to AI systems as or more capable than humans across most or all capability dimensions. In 2030, AI systems can operate with levels of autonomy and cognitive ability that match or surpass humans in cognitive tasks, autonomously working towards broad strategic goals that they can reflect upon and revise if circumstances change, while also collaborating with humans where necessary. AI-guided robots can handle complex tasks in dynamic real-world environments in many industries and roles, though they still largely lag humans in these roles unless developed specifically for that role.

The state of the evidence is insufficient to discount any of the scenarios outlined in this paper, or variations thereupon. The views of consulted experts aligned with this assessment. This suggests that AI progress by 2030 has a plausible range that includes both a plateau at approximately today's level of capabilities and rapid improvement that leads to AI systems which broadly surpass human capabilities.

Consulted experts expressed high uncertainty and low confidence in their ability to predict the rate of AI progress by 2030 and beyond. This reflects the extremely rapid rate of innovation in AI systems over recent

years combined with high uncertainty about the extent to which recent drivers of AI progress will continue to drive further progress.

1. Introduction and methodology

1.1. Understanding possible AI trajectories will enable governments to capture the benefits and prepare for potential impacts

AI has advanced rapidly in recent years. AI systems are now able to draft academic essays at the level of university students and solve coding problems at the level of human programmers (Yeadon et al., 2025[2]; Hou and Ji, 2024[3]). More advanced AI systems are routinely developed while governments, economies and societies are racing to keep up with the pace of change.

Governments would benefit from a better understanding of how AI could continue advancing. This understanding will help inform policy decisions that best capture the benefits of AI progress and prepare for its potential impacts.

To address this pressing policy need, the OECD has developed a set of expert and evidence informed scenarios for how AI could advance by 2030. While AI advances through 2030 cannot be predicted with certainty, governments can consider the range of plausible trajectories informed by the best available evidence and expert insights. This scenarios analysis aims to provide that baseline, outlining a range of plausible trajectories for AI progress by 2030.

1.2. Exploring trends and uncertainties to build four core scenarios for plausible AI trajectories through 2030

This analysis is supported by a combination of inputs:

a. **Review of relevant literature:** this paper draws on a wide range of cutting-edge research and evidence to inform its analysis.

b. **Interviews and review by leading AI experts:** this paper draws on inputs from experts with diverse backgrounds and perspectives. Experts interviewed or involved in reviewing this analysis are detailed in Annex A.

c. **Scenarios analysis using strategic foresight methods:** this paper employs strategic foresight methods to test assumptions and build scenarios about plausible AI futures. Strategic foresight methods used include trend analysis, horizon scanning, driver mapping and technology road mapping.

d. **Trend extrapolation to supplement the scenarios analysis:** this paper draws on existing data of historic trends in AI progress to extrapolate plausible rates of progress through to 2030. Rather than being the only method used to generate the scenarios, this trend analysis supplements the scenarios and helps to make them cohesive and concrete.

The scenarios explored in this paper are plausible but uncertain futures intended to inform policy discussions, not predictions. Given the uncertainty regarding future AI trajectories, probabilities are not

assigned to the different scenarios. The analysis of AI capabilities in this paper is based on information available up to October 2025.

2. AI progress trends and uncertainties

2.1. AI systems have demonstrated rapid progress on a wide range of benchmarks

Over recent decades, the performance of leading AI systems has improved quickly on a wide range of benchmarks and tests (see Annex B). On a benchmark of PhD level science questions, AI systems now outperform human experts, a rapid improvement from scoring only slightly better than chance in 2023 (Epoch AI, 2025[4]). In 2025, AI systems achieved gold medal level performance in the International Mathematical Olympiad, a prestigious competition for pre-university mathematicians (Kazemi et al., 2025[5]; Metz, 2025[6]). They also achieved gold-medal level at the International Collegiate Programming Contest World Finals, where top university teams compete globally to solve complex programming problems (Lin and Cheng, 2025[7]). AI systems continue to improve their ability to solve longer, more complex tasks autonomously, including tasks such as vision-guided computer use, software engineering, video interpretation, and simulated object manipulation (METR, 2025[8]). AI systems have also advanced rapidly in their multilingual abilities, achieving human parity in a benchmark of translation quality for widely spoken languages (Proietti, Perrelle and Navigli, 2025[9]). A new benchmark testing AI systems on precisely-specified digital tasks performed by workers from 44 occupations (ranging from industrial engineers to nurses) found that leading AI system's outputs matched or were preferred to human outputs 47.6% of the time by expert graders, indicating near parity with human performance on these tasks (Patwardhan et al., 2025[10])[1]. Benchmarks and tests such as these are imperfect, but they represent developers and experts' best efforts to quantitatively assess the abilities of different AI systems.

Despite these rapid gains, humans still outperform AI systems in important areas. AI systems lag in several areas such as continual learning, metacognition, agency, solving dynamic and real-world problems, generalising to solve novel problems, creativity, physical tasks and social interaction in dynamic social contexts (OECD, 2025[1]). Issues of robustness and hallucinations continue to substantially impact reliability (Song, Han and Goodman, 2025[11]). AI performance is also highly uneven across languages, with AI performance on reasoning tasks dropping substantially in low resource languages (Xuan et al., 2025[12]). For further discussion of trends in AI system capabilities, ongoing limitations, and limitations of AI benchmarks, see Annex B.

2.2. Key uncertainties about future trends in AI progress

2.2.1. The relationship between scaling of pretraining[2] and performance gains

In deep learning, the AI system learns patterns from data instead of being explicitly programmed. Deep learning models have "parameters", which are numbers used by the models to encode everything they learn from the data. The value of these parameters is set by "training" the model on a large amount of data. Through training, the model gradually updates the parameter values as it sees those data, so that

it gets better at whatever objective it is being trained for. This process of training requires computing power ("compute") to process the data and update the parameters. Compute is also required to use the model once it has been trained. This is known as "inference compute" or "test-time compute".

In recent decades, researchers identified that AI models became predictably more capable as the number of parameters, amount of training data and amount of compute used in training are increased. These observed relationships are known as "scaling laws" (Hestness et al., 2017[13]; Kaplan et al., 2020[14]; Lee et al., 2025[15]).

This scaling of AI models using more parameters, training compute and data has been the central driver of progress in frontier AI systems over the last decade, **with increased model scale strongly predicting performance on AI benchmarks** (Epoch AI, 2025[4]; Paglieri, Cupial and Piterbarg, 2024[16]; Owen, 2024[17])**.** Since 2010, the number of parameters in frontier AI models has increased by 2.4x per year, the amount of data used to train frontier models has increased by 2.6x per year and the amount of compute used to train frontier AI models has more than quadrupled each year (Epoch AI, 2025[18]).

However, these scaling laws are consistent trends observed from past data, not immutable rules. Theoretical explanations of scaling laws exist and provide partial justification for their robustness, but they do not guarantee that further scaling will continue to yield practically useful performance improvements (Bahri et al., 2024[19]; Brill, 2024[20]).

Continued scaling may continue delivering steady gains in AI systems' performance, in line with scaling laws or they might lessen or plateau (Dohmatob et al., 2024[21]; Chen et al., 2025[22]; Caballero et al., 2023[23])**.** Some experts argue that AI systems work primarily through sophisticated memorisation and interpolation, with their vast knowledge base concealing weaknesses in reasoning (Song, Han and Goodman, 2025[11]). If this is the primary driver of current AI capabilities, then AI systems could reach a capability ceiling above which more flexible and data-efficient reasoning would be required for cost-effective progress. However, this characterisation is disputed and remains an active area of research. Recent evidence suggests that advanced AI models perform key elements of reasoning to at least some extent, such as abstraction (learning reusable concepts from specific examples), compositionality (combining those concepts flexibly), and systematic reasoning (reasoning consistently about concepts according to rules) (Ni et al., 2025[24]; Prabhakar, Griffiths and McCoy, 2024[25]). Study of the internal workings of generative AI systems found that these systems develop emergent architectures that support abstract reasoning by generating new abstract concepts and using those concepts to inform predictions (Yang et al., 2025[26]; Du et al., 2025[27]). These findings suggest that apparent gains in reasoning performance may reflect more than memorisation. However, the robustness, generality, and cost-effectiveness of these reasoning capabilities remain uncertain, as models continue to show deficits in systematic reasoning tasks beyond those in their training data (Heyman and Zylberberg, 2025[28]; Khalid, Nourollah and Schockaert, 2025[29]). For further discussion of these uncertainties and the evidence gained from recent prominent models such as GPT-4.5 and GPT-5, see Annex B.

2.2.2. Gains from reinforcement learning for reasoning and scaling of inference compute

AI companies are increasingly focused on new approaches designed to produce "reasoning" in AI models. With the release of OpenAI's o1 series of models in 2024, foundation model developers started more deliberately scaling the amount of inference compute used by models during their operation (OpenAI, 2024[30]). This allowed models to think through problems step-by-step, iterating and verifying outputs during inference to enhance reasoning capabilities. This led to the creation of what developers call "reasoning models" (Wei et al., 2023[31]). Companies also began training their models to be better at this form of step-by-step reasoning. This step-by-step process requires more compute in the same way that stopping to think through a problem in detail requires more time and attention for a human than giving an answer off the top of their head. Developers use reinforcement learning to reward models for productive approaches

to reasoning, employing process supervision to reward correct reasoning steps rather than solely the final outcome. This involves using datasets of problems with known, easily verifiable answers to reward reasoning that arrives at correct solutions (Yue et al., 2025[32]).

Training models to reason using reinforcement learning marks a notable departure from the previous paradigm of training models to predict the next word from vast training datasets. Instead of training AI models to mimic text from the internet, reinforcement learning attempts to train AI models to engage in effective reasoning to produce correct answers. Using reinforcement learning to train models to reason has allowed AI developers to continue producing rapid progress in AI model performance (OpenAI, 2025[33]).

This progress in reasoning is not driven by reinforcement learning in isolation, but through an interplay with advances in AI system architectures, training and data quality (Liu et al., 2025[34])**.** Architectural innovations enable more efficient step-by-step reasoning, while improvements in training methods allow more to be learned from a given dataset. Developers have also invested in curating high quality datasets, including those showing step-by-step reasoning (DeepSeek, 2025[35]).

However, gains from reasoning appear to be less generalisable than gains from pretraining (Altman et al., 2025[36]; Alam and Rastogi, 2025[37])**.** It is currently uncertain how well these approaches will generalise to different forms of reasoning and how successful AI developers will be at finding or producing data to train different forms of reasoning.

Progress is likely to be most rapid in the types of tasks for which it is easiest to conduct reinforcement learning from verifiably correct answers, such as mathematics. If rapid gains in reasoning capabilities continue with further reasoning training and generalise well, then AI systems could develop more advanced capabilities. If this is not the case, AI systems could remain limited to narrower forms of reasoning for which solutions are easy to verify and data are abundant or easy to produce, such as mathematics and structured coding tasks.

2.2.3. Progress in memory and continual learning

Developers may succeed in developing AI systems with more reliable memory and learning, or this may remain a bottleneck for AI performance. Current progress in AI memory largely relies on increasingly large context windows along with noting and retrieval of key information (see Annex B). If progress is made on continual learning, AI systems could move beyond these limited approaches. Alternatively, these limited approaches might continue being iterated and improved to gradually approximate continual learning. For instance, new techniques may make AI systems better able to identify and use only the most relevant and important information within their large context windows.

2.2.4. Progress in physical capabilities

If current trends continue, progress in physical capabilities will remain slower than for purely cognitive tasks, but there is potential for breakthroughs or the emergence of new trends enabling more rapid progress. Improvements in visual and multimodal perception have significantly enhanced robots' ability to interpret and reason about their environments (Zhao, Gangaraju and Yuan, 2025[38]). However, translating these gains in perception into robust, general physical capabilities remains challenging. Current progress in the physical capabilities of robots is limited by a number of challenges, such as insufficient high-quality data, weak robustness of capabilities to new environments or tasks, limited causal reasoning, and weaknesses in long-sequence tasks (Li et al., 2025[39]; Liu et al., 2024[40]).

Specialised AI systems in constrained settings such as prosthetics, assistive technologies and specialised industrial systems have seen significant progress, while general purpose physical capabilities continue to face substantial limitations (Sarkar and Alqasemi, 2025[41]; Urrea and Kern,

2025[42]). Multi-modal large models and world models may offer approaches to overcome some of these limitations due to their strong combination of perception, interaction and reasoning capabilities. This makes them a promising architecture for physically embodied AI agents (Liu et al., 2024[40]). Robotic components are also becoming cheaper, enabling more widespread experimentation and data generation (Shaw, Agarwal and Pathak, 2023[43]).

2.2.5. Progress on robust agentic behaviour and metacognition

Continued investments may produce more robust metacognition and agentic behaviour, or this may prove difficult to achieve. AI systems are currently substantially less effective than humans at monitoring and correcting their own reasoning while pursuing goals, limiting their ability to autonomously complete longer and more complex tasks (OECD, 2025[1]). However, AI developers are currently investing substantial effort into producing more capable AI agents (OpenAI, 2025[44]). Investments in reinforcement learning or other forms of training to improve the ability of AI agents to break down tasks, follow multi-step plans, use tools, and calibrate their uncertainty could produce more competent AI agents, continuing current trends in progress on agentic tasks (see Section 2.2.2). Developers are also leveraging multi-agent architectures to build more competent agentic systems, where specialised sub-systems or "agents" collaborate and cross-verify outputs to enhance robustness (Tran et al., 2025[45]).

2.2.6. Progress on creativity and the ability to solve novel problems

AI systems lag humans in creativity and their ability to solve problems that do not align with their training data (see Annex B discussion of creativity). This weakness stems from AI systems' reliance upon their training data to infer statistical associations. However, new approaches to developing AI systems are emphasising reinforcement learning for more advanced reasoning capabilities (OpenAI, 2024[30]). These approaches have been observed to allow AI systems to creatively solve problems in ways that humans have never considered. For example, DeepMind's AlphaGo which was trained via reinforcement learning used unconventional moves that commentators described as creative and unlike moves from human go players (Metz, 2016[46]). Other approaches such as evolutionary agents can produce novel and creative solutions on certain tasks (Google Deepmind, 2025[47]).

The integration of traditional generative AI models trained on vast datasets with approaches such as reinforcement learning could enable greater creativity and novel problem solving. It is also possible that greater creativity could emerge as AI systems are further scaled or if there are further algorithmic breakthroughs.

2.3. Key uncertainties about future AI inputs

2.3.1. Scaling compute and data inputs

Most progress in generative AI performance over 2012-2023 came from increasing the size, training compute and training data used to develop models (Ho et al., 2024[48])**.** In 2024 and 2025, continued progress has also been driven by training models to reason and increasing the compute used by AI models after deployment to allow models to reason for longer (see Sections 2.2.1 and 2.2.2).

Continued progress using each of these approaches will require further scaling the compute and data available to train AI systems and the compute available to run AI systems. Continued scaling would require increasingly expensive, data intensive, compute intensive, water intensive and power intensive training runs. The cost of the compute used to train frontier models rose by 2.8x per year from 2015 to 2024 (Epoch AI, 2025[18]). The power used to train frontier models increased 2.3x per year from 2011 to 2025 (Epoch AI, 2025[18]).

Estimates of the ability to scale the amount of compute and data used to train AI models suggests that continued scaling in line with current trends is possible through to 2030 (Sevilla et al., 2024[49])**.** This would enable training runs using 10,000 times greater compute than that used to train OpenAI's GPT-4. If scaling continued at its current pace, power constraints are predicted to begin limiting the rate at which AI training runs can be scaled further around 2030, with other constraints potentially posing barriers at a similar or later date. Potential limitations to compute scaling due to power requirements and water consumption remain critical uncertainties regarding future progress. Scaling of datasets is likely to continue to be possible during this period, in part facilitated by "self-play" techniques where models generate and verify their own synthetic training data (Sevilla et al., 2024[49]). For further discussion of current trends and potential future barriers to scaling, see Annex C.

Despite the estimates that indicate continued scaling is possible, uncertainty remains. Training runs of this size are unprecedented, and unexpected barriers could prevent them from being achieved. Training runs of this size would also require high levels of investment, which will be dependent upon continued economic returns to scaling AI models. This is uncertain if scaling doesn't provide sufficient capability gains or capability gains are not sufficiently valued by consumers.

2.3.2. Algorithmic efficiency gains

If the rate of algorithmic innovation since 2012 continues, it could drive substantial performance gains. From 2012 to 2023 the compute budget needed to achieve a fixed AI performance level has halved roughly every 8 months, faster than Moore's law (Ho et al., 2024[48]). However, the conditions necessary for this continued pace are uncertain. If the most impactful and readily identifiable algorithmic advancements have already been implemented, algorithmic progress could slow. Alternatively, new transformative innovations could allow continued or even more rapid capabilities growth. It is also uncertain the extent to which rapid algorithmic advancement is dependent upon continued compute scaling, meaning future algorithmic innovation could be impacted by the extent of compute scaling that is achieved (Josephson, 2025[50]; Barnett, 2025[51]). For further details on trends and uncertainties in algorithmic innovation, see Annex C.

2.3.3. The use of AI systems in AI development

AI performance in software engineering is advancing rapidly. Most software developers use AI tools daily and believe AI will increase the quality of their code (Stackoverflow, 2025[52]; GitHub, 2025[53]). In December 2024, AI wrote an estimated 29% of the Python code produced by US programmers, and this proportion has been increasing rapidly since 2021 (Daniotti et al., 2026[54]). AI systems increasingly support a wide range of coding tasks such as debugging, restructuring and testing code, including some tasks which require reasoning across multiple files or larger codebases (Otten et al., 2025[55]). However, robustness and security issues, such as AI producing insecure code, hallucinating, or exposing confidential information, continue to limit full automation of coding tasks (Cotroneo, Cristina and Liguori, 2025[56]; Al-Maamari, 2025[57]).

Accelerating the development of future AI systems using AI is a core part of developers' strategy (OpenAI, 2023[58])**.** Staff at leading AI companies already use AI coding assistants to support model development, experimentation and infrastructure engineering, and AI systems have already autonomously produced improved algorithms to support data centre efficiency, chip designs and AI training processes (Anthropic, 2025[59]; Google Deepmind, 2025[47]).

However, the extent to which AI supports software engineering productivity now and in the future is uncertain. Randomised control trials at Microsoft, Accenture and another Fortune 100 company found that AI coding assistants increased the rate at which software developers completed tasks by 26% (Cui et al., 2025[60]). Another randomised control trial found that coding assistants improved task completion

time by 30% while maintaining or improving code quality and maintainability (Borg et al., 2025[61]). However, a different randomised control trial found that AI tools decreased the productivity of experienced AI developers in their areas of expertise, slowing them by approximately 20%, despite the developers believing it had sped their work (Becker et al., 2025[62]). The reason for these divergent results is an area of active research. The impacts of AI systems on progress in AI could be limited or transformative, depending on how capable AI systems become at software engineering and where bottlenecks exist in the development of more advanced AI.

2.3.4. Other social, economic and institutional factors influencing AI progress

Broader contextual factors, such as regulation, policy, economic conditions, AI diffusion and public opinion could all impact the pace and direction of AI progress. Similarly, the ownership, distribution and governance structures of AI development and inputs such as training data could substantially impact AI trajectories. Economic drivers of AI progress are a key uncertainty, with returns to AI progress dependent upon adoption and associated productivity gains or other benefits. A productivity lag could weaken future investment, while rapid productivity or other benefits from adoption could accelerate investment. Current concentration of capital, computing power and talent in a small number of frontier AI companies could similarly speed or slow AI progress, depending on the relative benefits of consolidation versus diversity in AI research and development efforts. Market and geopolitical fragmentation could also impact AI progress, potentially limiting access to computing power, data and human capital or triggering substantial national-level investments. Both the direction and magnitude of these influences upon AI development through 2030 is uncertain. These factors are explored in more detail in *Futures of Global AI Governance: Co-Creating an Approach for Transforming Economies and Societies* (OECD, 2024[63]). Many of these uncertainties are not unique to AI, with technologies such as quantum technologies and biotechnology also facing high degrees of uncertainty. Nonetheless, these uncertainties could have significant impacts upon the rate of AI progress (Robinson and Nadal, 2025[64]; OECD, 2025[65]).

Scaling of AI systems is dependent upon access to energy to power training and inference and water for cooling. Policy decisions will shape access to energy and water resources for AI developers, and will involve trade-offs with other uses and sustainability considerations. Data centres are already impacting local water supplies in some cases, and a data centre project in Chile was temporarily halted until concerns about water use were addressed (Associated Press, 2024[66]; Klienman and Wheeler, 2025[67]). However, new data centre designs are being adopted that remove the need for fresh water supply, which may mitigate this constraint where the new designs can be adopted (Solomon, 2024[68]). Similarly, data centres are predicted to account for 20% of the growth in electricity demand in the International Energy Agency's advanced economies regional grouping through to 2030, with AI as a dominant driver (International Energy Agency, 2025[69]). This follows decades of stagnant electricity demand in advanced economies, meaning advanced economies will face policy decisions to address rising demand.

3. Scenarios

3.1. Using the OECD's AI Capability Indicators (beta) to define AI capability categories

This paper draws on the OECD's AI Capability Indicators (OECD, 2025[1]) to define AI capability categories. The indicators are designed to assess and compare AI advancements against human abilities. They were developed and reviewed by a network of AI researchers, psychologists, and other experts. They provide a framework for policymakers to understand AI's potential impacts on education, work, and other areas of life.

The nine indicators cover a range of policy-relevant human abilities:

- **Language:** the ability to understand, interpret and generate human language. The language capability is assessed along six dimensions: linguistic (semantics/grammar, tone and emotional expression), modality (verbal vs text, understanding vs generation), number of languages and three dimensions covering the language-related aspects of knowledge, reasoning and learning.
- **Social interaction:** the ability to perceive, interpret and appropriately respond to social cues in dynamic interpersonal contexts, including extended or embodied interactions with multiple individuals. The social interaction capability is assessed along seven dimensions: embodiment, social memory, identity, social communication, affective skills, social perception, and social problem solving.
- **Problem solving:** the ability to integrate qualitative, quantitative and logical information through multi-step reasoning, including analysis, prediction, explanation and counterfactual thinking. The problem-solving capability is assessed along four dimensions: the types of solutions required, the range of alternatives considered, the complexity of expert knowledge, and the complexity of model formulation and interpretation.
- **Creativity:** the ability to produce valuable, novel, transformative and surprising outputs with intentionality and adaptability and to self-assess creative merit. The creativity capability is assessed along five dimensions: value, novelty, adaptability, intentionality and self-assessment.
- **Metacognition and critical thinking:** the ability of an AI system to evaluate its own reasoning, calibrate confidence and identify relevant information in complex tasks. This capability is assessed along three dimensions: critical thinking (such as assessing strategy and self-monitoring), confidence calibration (an AI system knowing what it does and does not know) and identifying which information is relevant.
- **Knowledge, learning and memory:** the ability to structure information as knowledge, acquire it via learning, and store and retrieve it via memory. This capability is assessed along three dimensions: kinds of knowledge, learning mechanisms and memory processes.
- **Vision:** the ability to interpret visual scenes in their full complexity, with a wide range of visual conditions and environments. This capability is assessed along four dimensions: breadth and variability of the object in focus, variation in background and visual environment, learning and task diversity.
- **Physical manipulation:** the ability to interact with physical objects in the environment, which includes the physical movements themselves; the necessary perception, including tactile, visual or other sensors, to provide feedback; and cognition to plan and adjust the movements. This capability is assessed along four dimensions: the range/type of movement, object characteristics,

environment and time pressure.

- **Robotic intelligence:** the ability to act as an autonomous agent in a natural environment, involving the co-ordination of the full range of human abilities. This capability is assessed along six dimensions: environment, task, abstraction, human interaction, uncertainty and ethics.

These indicators are presented in scales of five levels, with higher levels representing more advanced capabilities. A level of 5 indicates performance equivalent to human-level. When the indicators were published in mid-2025, reflecting AI performance in late 2024, the capabilities of the most advanced AI systems were assessed by this framework as reaching levels 2 or 3 across all nine capabilities (Table 1).

Table 1. AI performance relative to OECD AI capability indicators, reflecting AI performance in late 2024 (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic Intelligence
3	2	2	3	2	3	3	2	2

Source: (OECD, 2025[1])

These indicators are used in the scenarios to highlight the approximate capability levels that AI systems might achieve per indicator in each scenario.

3.2. Building on these indicators to explore scenarios for AI progress in 2030

Based on current trends in AI, inputs to AI, key uncertainties and potential future disruptions, this section outlines four plausible scenarios for AI progress by 2030. These are:

a. **Progress Stalls:** a scenario in which AI progress largely halts and the performance of AI systems plateaus. After 2025, gains within existing machine learning approaches begin to hit fundamental limits and significant new innovations (akin to the transformer architecture) are not developed.
b. **Progress Slows:** a scenario in which incremental gains within existing approaches to AI development deliver continued but slower progress. After 2025, deep learning approaches begin to reach maturity and progress slows as low hanging fruit have been picked.
c. **Progress Continues:** a scenario in which continued rapid progress occurs. This could occur due to continued investment in larger AI training runs or continued algorithmic innovations as impactful as those seen over the past decade.
d. **Progress Accelerates:** a scenario in which dramatic progress leads to AI systems as or more capable than humans across most or all capability dimensions. These rapid gains could be driven by a combination of continued gains from investment in larger AI training runs, further algorithmic breakthroughs, and increasing contributions from AI systems themselves as software engineering assistants.

AI systems today have an uneven set of capabilities, lagging in some areas (such as physical capabilities) and excelling in others (such as language capabilities). This unevenness is also present within capability categories. For example, in the category of knowledge, learning and memory, leading AI systems excel at

processing and retrieving large volumes of data, but largely lack continual learning capabilities (OECD, 2025[1]). The primary scenarios in this paper (Progress Stalls, Progress Slows, Progress Continues and Progress Accelerates) assume that the relative rates of progress on different AI capabilities remain roughly in line with their current trends[3]. However, it is possible that rates of progress on different AI capabilities diverge from current trends, producing sets of uneven capabilities, which could lead to AI systems with different capability profiles from those discussed in this paper's primary scenarios. To address this uncertainty, each scenario also includes variations illustrating potential deviations from the scenarios.

3.3. Scenario 1: Progress Stalls

Box 1. Scenario 1: Progress Stalls

Summary: In this scenario, progress in frontier AI systems plateaus soon after 2025 and their capabilities are largely unchanged by 2030. The most capable AI systems in mid-2030 closely resemble the most capable systems of 2025, though they have been refined for user-friendliness and integration into applications. Diffusion, adoption and application development for existing capabilities continues to occur. Multimodal AI systems can process text, speech, images and video to perform a wide range of short, clearly scoped tasks. AI systems typically rely on substantial support from humans to complete tasks, such as detailed prompting, review and provision of the relevant context. They excel in terms of their knowledge base and ability to answer knowledge-based questions, though issues with hallucinations still impact reliability. They are capable at some university level mathematics and other forms of structured reasoning problems, but continue to struggle with dynamic, novel or real-world problems. They largely lack robust abilities to learn new skills or form memories. Memories are limited to the equivalent of keeping notes and snapshots for later reference combined with an ability to consider a large amount of context at once, lacking the flexibility and efficiency of human memory and recall abilities. AI systems still struggle to maintain coherent thinking for longer or more complex tasks, limiting their ability to act as independent agents. AI systems remain far from human equivalence in their ability to interact with the physical world and handle physical tasks in complex and dynamic environments, handling only simple tasks in heavily controlled environments or more complex environments after extensive targeted training and testing. AI systems can analyse social interactions in text format, but they lack flexible and responsive social skills required to integrate into human social environments, such as smoothly contributing to a multi-person conversation, with humans instead needing to adapt to AI systems in order to interact with them.

How we get here: After 2025, gains within existing approaches for developing frontier AI models hit fundamental limits. These could include a stalled relationship between increased model size or reasoning training and capability gains, limits on the ability to continue scaling the compute and data used in training, or a significant drop-off in AI investment. The approaches of frontier model developers hit hard limitations in the ability to deliver continual learning, metacognition and agency, problem solving, creativity, physical manipulation and robotic intelligence, and social interaction, with existing training paradigms hitting a ceiling in what they can deliver. Transformative new innovations (like the transformer architecture or reasoning models in the past several years) do not materialise to prevent a plateau. Rapid improvements in AI cease to occur, and AI progress plateaus, resulting in models with roughly similar capabilities to those available in 2025. This scenario could also arise if investment in AI decreases drastically, public caution stalls adoption and development, or policy deliberately or unintentionally stalls development. Alternatively, stalling progress due to technical limitations could coincide with high investment and broad societal acceptance.

Historical analogue for this scenario: The speed of passenger aircraft rose rapidly from 1930 to 1960, before plateauing at approximately 500 knots. Physical limitations such as the sound barrier and rapid increases in drag above this speed made faster aircraft impractical for passenger travel (IPCC, 1999[70]).

Indicative capabilities from trend extrapolation to mid-2030:
The length of tasks (measured in the time it takes a human expert to complete them) that an AI system can finish successfully

more than 50% of the time.
These indicative estimates are produced by taking the estimated length of tasks AI systems can complete today from benchmarking studies and assuming AI systems will plateau at double this current task horizon (METR, 2025[81]). *See Annex D for further details.*

Scientific reasoning (answering PhD-level scientific reasoning questions):	1 days	Computer use (using a computer to achieve a specific goal):	3 minutes
Mathematical reasoning (solving challenging math problems):	3 hours	Web navigation (navigating real websites to achieve a specific goal):	4 minutes
Software engineering (completing coding tasks autonomously):	4 hours	Simulated robotics (controlling a simulated robotic arm to achieve a specific goal):	3 minutes
Autonomous driving (driving duration without need for human intervention):	3 hours		

Table 2. Scenario 1: AI Progress Stalls – capability indicator scores (1-5 scale)

	Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic Intelligence
2030 score	3	2	2	3	2	3	3	2	2
Change from 2025	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0

These indicator scores are hypothetical, intended to communicate the highest capability levels that AI systems might achieve on the OECD's AI Capability Indicators in this scenario. The indicators are on a scale from 1-5 with a level of 5 corresponding to roughly human-level capabilities (see Section 3.1).

3.3.1. Reasoning supporting the plausibility of this scenario

Capability gains from scaling AI models may not continue. There are some indications that simply increasing the size, compute and data used to train AI models is already producing weaker capability gains than developers anticipated (see Annex B). There may also be fundamental limits to the capabilities AI systems can acquire through existing AI development paradigms or limitations to the physical infrastructure, investment, and resources needed to scale AI systems.

Capability gains from training models to reason may fail to generalise well beyond domains such as mathematics and coding, and the ability to scale these approaches may become substantially more limited beyond 2026 (see Section 2.2.2).

Developers may reach fundamental limits on their ability to scale compute and data inputs sooner than expected (see Section 2.3.1).

Further algorithmic innovations may not emerge, mimicking past “winters” in AI development (see Annex C).

AI systems may fail to deliver meaningful gains in the productivity of software developers and societies more broadly (see Section 2.3.3).

3.3.2. Reasoning against the plausibility of this scenario

It is likely that some algorithmic innovations that advance AI capabilities will occur in the period from 2025 to 2030, based on past trends in algorithmic innovation (see section 2.3.2 and Annex C).

Even if capability gains from continued scaling and reasoning training are less than expected, they are still likely to deliver at least some incremental improvement (see sections 2.2.1 and 2.2.2).

3.4. Scenario 1: Potential variations

Examples of variations for a scenario in which Progress Stalls could include:

a. AI as a Narrow Tool
b. Simple AI agents

Box 2. Variant A: AI as a Narrow Tool

This is a potential variant of the Progress Stalls scenario

AI systems develop more advanced problem solving for limited tasks, such as coding and mathematics, but this fails to generalise well beyond their training data. Systems are limited by metacognition and agency, so they are unable to chain tasks to deliver higher level objectives over longer timescales. This could occur if developers struggle to produce solutions to AI system's ability to continually learn and tendency to veer off track and lose coherence on longer tasks, but succeed in training AI systems to become more effective problem solvers on narrow sets of tasks through reinforcement learning for specific forms of reasoning. Examples of narrow tools that could advance in this scenario include coding support tools, mathematical theorem proving systems, decision support systems in narrow domains (such as supporting medical diagnoses or legal processes), or design tools for materials, biology or mechanical systems.

Table 3. Variant A: AI as a narrow tool – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
3	2	3	3	2	3	3	2	2

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

Box 3. Variant B: Simple AI Agents

This is a potential variant of the Progress Stalls scenario

Progress largely stalls on creating more generally capable AI systems, but the efforts of AI developers to build more effective AI agents have some success. These agents have more advanced metacognition that allows them to more autonomously identify relevant details, assess their strategies and assess their progress on tasks. This enables them to autonomously perform simple computer-based tasks on behalf of users with acceptable levels of reliability, as well as more complex tasks in fields with rich sources of training data such as software engineering. These AI systems remain limited in their ability to exhibit creativity and generate novel thinking to solve problems, and struggle with social or real-world problem solving. This could arise through devoted training efforts from AI developers to better elicit the current capabilities of large language models to critically assess and monitor their performance on agentic tasks.

Table 4. Variant B: Simple AI Agents – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
3	2	2	3	3	3	3	2	2

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

3.5. Scenario 2: Progress Slows

Box 4. Scenario 2: Progress Slows

Summary: In mid-2030, AI systems are markedly more capable than systems today, despite slower capability gains after 2025. AI systems have an exceptional knowledge base and can answer questions effectively on any expert-level topic. Hallucinations are still occasionally present for knowledge-based questions, but significantly reduced, meaning answers are typically reliable. AI systems typically rely on humans to develop clearly scoped tasks, review important decisions or actions, and provide detailed guidance and relevant context. AI systems are capable of researcher-level structured reasoning in fields such as mathematics and science. AI systems can maintain coherent thinking and error-correct to perform longer or more complex tasks. They are able to autonomously and quickly perform a range of agentic tasks that would take a human hours to days, meaning they are often able to act as useful assistants for tasks that require them to use a computer, navigate the web or undertake limited interaction with people or services on behalf of the user. AI systems can complete these tasks substantially more quickly than a human in most cases, and while not perfectly reliable are often competitive with human levels of reliability on many tasks. AI systems have improved memory and can distil down key facts that they need to recall and effectively access them when needed. However, AI systems remain relatively weak at continual learning, meaning they struggle to learn new skills or approaches to problems after deployment. This weakness can be partially addressed through fine-tuning for particular tasks or industries, which has been made simpler and more accessible. AI systems remain far below human equivalence in their ability to interact with the physical world and handle physical tasks in complex and dynamic environments. They can handle complex longer physical tasks in simulated environments, but struggle to transition this to the complexities of real-world environments. As a result, robotic capabilities remain mostly limited to controlled environments (such as laboratories or warehouses), but can include increasingly complex and dynamic tasks within these environments. AI systems continue to lack the flexible and responsive social skills required to integrate seamlessly into human social environments, but can now engage coherently in simple interactions with multiple stakeholders or more fluent social interactions in one-on-one settings. The enhanced memory of AI systems helps them maintain a somewhat more coherent persona across interactions and develop a basic understanding of people and their motivations to guide their interactions with different people.

How we get here: Progress in AI continues, but slows substantially. Deep learning approaches reach maturity and low hanging fruit have been picked, despite early rapid growth. The approaches of frontier model developers struggle to overcome limitations in continual learning, metacognition and agency, problem solving, creativity, physical tasks and robotic intelligence, and social interaction, with existing training paradigms providing imperfect solutions and algorithmic innovations making only gradual progress. Increased model size and reasoning training continue to produce performance gains, but at a lesser rate than was hoped by AI developers. This limits the ability to continue scaling the compute and data used in training as investors see lower returns from continued investments. Unforeseen bottlenecks in investment, infrastructure, natural resources, data supply and energy limit the ability to scale compute and data. This scenario could also arise if investment in AI declines, public caution slows adoption and development, or policy deliberately or unintentionally slows development. Alternatively, slower progress due to technical limitations could coincide with high investment and broad societal acceptance.

Historical analogue for this scenario: Antibiotic discovery, which experienced a "golden era" of rapid new antibiotics discoveries from the 1940s to 1960s, then experienced slowing in discoveries as the low-hanging fruit from existing methods for antibiotic discovery were exhausted (Lewis, 2020[71]).

Indicative capabilities from trend extrapolation to mid-2030:
The length of tasks (measured in the time it takes a human expert to complete them) that an AI system can successfully finish more than 50% of the time.
These indicative estimates are produced by taking the estimated length of tasks AI systems can complete today and the

rate at which this task horizon is doubling from benchmarking studies (METR, 2025[8]). *This scenario assumes the doubling rate in task lengths will slow substantially, with each doubling taking 30% longer. See Annex D for further details.*

Scientific reasoning (answering PhD-level scientific reasoning questions):	>1 month	Computer use (using a computer to achieve a specific goal):	3 hours
Mathematical reasoning (solving challenging math problems):	17 days	Web navigation (navigating real websites to achieve a specific goal):	1 hour
Software engineering (completing coding tasks autonomously):	2 days	Simulated robotics (controlling a simulated robotic arm to achieve a specific goal):	1 hour
Autonomous driving (driving duration without need for human intervention):	10 hours		

Table 5. Scenario 2: Progress Slows scenario capabilities – capability indicator scores (1-5 scale)

	Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
2030 score	4	3	3	3	3	3	3	2	2
Change from 2025	+1	+1	+1	<1	+1	<1	<1	<1	<1

These indicator scores are hypothetical, intended to communicate the capability levels that AI systems might achieve on the OECD's AI Capability Indicators in this scenario. The indicators are on a scale from 1-5 with a level of 5 corresponding to roughly human-level capabilities (see Section 3.1).

3.5.1. Reasoning supporting the plausibility of this scenario

Each generation of models from OpenAI (GPT-2, GPT-3, GPT-4 etc.) uses approximately one hundred times more compute than the previous version (Altman et al., 2025[36]), making continued scaling highly cost-intensive and reliant upon substantial continued investment.

Scaling the size of AI models may continue to produce capability gains, but these gains could be weaker than past gains (see Annex B).

Training models to reason may produce only brittle forms of reasoning that fail to generalise well, slowing progress (see Section 2.2.20).

Developers may face increasing costs and operational constraints on their ability to rapidly scale compute and data inputs (see Section 2.3.1). For example, the growing prevalence of AI-generated data on the internet could negatively affect the ability to train future models on data scraped from the web (Shumailov et al., 2024[72]).

Algorithmic innovations may not continue to be identified at the rate they have over recent years if current paradigms are reaching maturity and new paradigms do not emerge (see Section 2.3.2).

AI systems may deliver limited gains in the productivity of software developers (see Section 2.3.3).

3.5.2. Reasoning against the plausibility of this scenario

Even if gains from reasoning training are brittle, AI developers are investing substantial resources in specialised training to make AI systems more capable at reasoning, social interaction and agentic tasks. This could drive rapid progress despite limitations.

Based on past rates of algorithmic innovation, it may be reasonable to expect additional transformative innovations on the level of the transformer architecture over the next five years. From 2012-2023 the compute budget needed to achieve a fixed AI capability level halved roughly every 8 months, faster than Moore's law, suggesting substantial rates of algorithmic innovation (Ho et al., 2024[48]).

The largest US technology companies are investing heavily in constructing data centres to enable further AI development and use, with the five highest-spending US enterprises projected to spend USD 736 billion in capital expenditure in 2025 and 2026 (Goldman Sachs, 2025[73]). This suggests an expectation that further scaling of AI systems will deliver returns, though it could also point to increased compute demands for inference to run AI models to serve customers.

3.6. Scenario 2: Potential variations

For a Progress Slows scenario, potential variations where rates of progress across different capabilities diverge from their current trends could include:

a. Simple Robots
b. Socially-Limited AI

Box 5. Variant C: Simple Robots

This is a potential variant of the Progress Slows scenario

New innovations in AI and robotics enable AI systems to achieve some of the flexible and adaptable performance on robotic tasks that language models currently demonstrate on language tasks. This enables simple robotic AI systems to be able to execute multi-step tasks in environments with dynamic changes such as changes in lighting, weather, clutter, distractions and the types of objects present.

Table 6. Variant C: Simple Robots – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
4	3	3	3	3	3	3	3	3

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

Box 6. Variant D: Socially-Limited AI

This is a potential variant of the Progress Slows scenario

AI systems may continue to become more capable at analysing and generating responses to described social situations or problems in language-based formats (such as text conversations). However, they may continue to struggle with social capabilities that involve interactions between multiple individuals, social memory, a consistent persona that is robust to adversarial or vulnerable actors and understanding of sensory inputs and physical environments. This could arise if AI systems struggle to data-efficiently learn human social information on the go, making them substantially less socially capable. Dynamic and real-world social interaction, where AI systems must perceive, interpret and respond in real-time to multimodal social cues may require the integration of sensory input, memory and world models to guide social interactions. Embodied social interactions are likely to be particularly challenging, requiring a complex combination of sensing, processing and operation of robotic components.

Table 7. Scenario Variant D: Socially-Limited AI – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
4	2	3	3	3	3	3	2	2

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

3.7. Scenario 3: Progress Continues

Box 7. Scenario 3: Progress Continues

Summary: In mid-2030, AI systems are substantially more capable than systems in 2025. Progress from 2025 to 2030 is as remarkable as that from 2020 to 2025. AI systems have a near-comprehensive knowledge base and can accurately answer questions on any expert-level topic. AI systems typically rely upon humans to provide high level directions and bounds for their behaviour, but can often operate with high autonomy to act within these bounds towards a given objective. AI systems outclass experts and professionals in many forms of structured reasoning in a broad range of fields. AI systems can maintain coherent thinking and error-correct to perform many professional tasks that would take a human roughly a month, such as executing a software engineering project, provided these tasks can be completed in a purely digital environment. AI systems can complete these tasks substantially more quickly than a human and with high reliability. They can ably navigate digital tools and environments autonomously, enabling them to behave with high autonomy where needed for clearly scoped roles. These abilities are supported by substantially improved abilities to learn and remember information. AI systems are effective at recording and retrieving important information across time. They lack fully flexible and automatic continuous learning, but can approximate this reasonably well through a combination of very large context windows, databases of key facts, and regular and automated cycles of fine-tuning and retraining. This produces AI systems that can "learn on the job" to some extent. AI systems have made substantial progress in their ability to guide robots and handle physical tasks in complex and dynamic environments. AI-guided robots can handle some complex tasks in dynamic real-world environments beyond factories and warehouses, including some tasks that require interaction with humans and adjustment to changing conditions. AI systems meet basic thresholds for the social skills required to integrate into human social environments, meaning they can coherently engage in diverse interactions with multiple stakeholders. Limitations remain for AI systems operating in real-world and/or advanced social environments, with AI systems still showing notably weaker problem-solving abilities in these environments than purely digital environments. The enhanced learning abilities of AI systems help them maintain a coherent identity across interactions and develop a detailed understanding of multiple people and their motivations to guide their interactions with different individuals, allowing them to effectively manage persistent social relationships.

How we get here: Rapid progress in AI continues uninterrupted through to 2030. This could occur if gains from scaling AI models and reasoning training continue, or new algorithmic innovations drive continued growth in line with past trends. Scaling of compute and data continue in line with current trends and do not hit substantial limits before 2030, matching current estimates of the room for continued growth. The approaches of frontier model developers succeed at overcoming many limitations in continual learning, metacognition and agency, problem solving, creativity, physical tasks and robotic intelligence, and social interaction. This occurs either through iteration and extension of existing approaches or novel breakthroughs. This scenario could also arise if investment in AI continues, wide public adoption rates continue, or policy deliberately or unintentionally supports continued development. Alternatively, rapid progress due to technical breakthroughs could also potentially coincide with slower public adoption and declining societal acceptance.

Historical analogue for this scenario: Moore's law, where computer chips doubled in computing power roughly every two years over five decades (Roser, Ritchie and Mathieu, 2023[74]).

Indicative capabilities from trend extrapolation to mid-2030:
The length of tasks (measured in the time it takes a human expert to complete them) that an AI system can successfully finish more than 50% of the time.
These indicative estimates are produced by taking the estimated length of tasks AI systems can complete today and the rate at which this task horizon is doubling from benchmarking studies (METR, 2025[8]). *This scenario assumes the doubling rate in task lengths will continue at current rates. See Annex D for further details.*

Scientific reasoning (answering PhD-level scientific reasoning questions):	1 month	Computer use (using a computer to achieve a specific goal):	>1 month

Mathematical reasoning (solving challenging math problems):	>1 month	Web navigation (navigating real websites to achieve a specific goal):	14 days
Software engineering (completing coding tasks autonomously):	24 days	Simulated robotics (controlling a simulated robotic arm to achieve a specific goal):	2 days
Autonomous driving (driving duration without need for human intervention):	1 day		

Table 8. Scenario 3: Progress Continues scenario capabilities – capability indicator scores (1-5 scale)

	Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
2030 score	4	3	4	4	4	4	4	3	3
Change from 2025	+1	+1	+2	+1	+2	+1	+1	+1	+1

These indicator scores are hypothetical, intended to communicate the capability levels that AI systems might achieve on the OECD's AI Capability Indicators in this scenario. The indicators are on a scale from 1-5 with a level of 5 corresponding to roughly human-level capabilities (see Section 3.1).

3.7.1. Reasoning supporting the plausibility of this scenario

Current analysis suggests it should be possible to continue scaling AI models in line with current trends (see Section 2.3.1).

Scaling the size of AI models has produced reliable capability gains over many orders of magnitude. GPT 4.5 and GPT-5 both achieved substantially higher scores on a range of benchmarks than earlier models (see Annex B).

Training models to reason may generalise reasonably well, continuing to produce more capable models as it is more extensively used (see Section 2.2.20).

Algorithmic innovations could continue to be identified at the rate they have over recent years. Even if scaling of learning from massive datasets has reached maturity, approaches to improving reasoning, tool use and memory could see rapid progress as they receive more focus from AI developers (see Annex C).

3.7.2. Reasoning against the plausibility of this scenario

Current gaps in continual learning, metacognition and agency, problem solving and creativity may prove difficult to resolve even with novel methods (see Sections 2.2.3, 2.2.5 and 2.2.6).

Existing progress has relied upon rapid scaling of compute used in AI training, but the gains from this scaling may not continue (see Annex B).

3.8. Scenario 3: Potential variations

For a Progress Continues scenario, potential variations where rates of progress across different capabilities diverge from their current trends could include:

a. Forgetful AI
b. Digital-only AI

Box 8. Variant E: Forgetful AI[4]

This is a potential variant of the Progress Continues scenario

AI systems are substantially more capable than systems today thanks to breakthroughs enabling improved metacognition, creativity and vision, but are limited by slow advances in memory and learning. Extension of current approaches such as larger context windows, databases of key facts, and regular retraining and fine-tuning fail to effectively approximate flexible continual learning. AI systems struggle to efficiently access and use information from within their large context windows and databases, resulting in overall weak performance on memory and learning. This weakness in memory and learning hampers natural social interactions, which depend on social memory as well as more advanced problem solving and language capabilities that depend upon contextual memory. AI systems in this scenario can effectively undertake multi-step tasks and produce useful creative outputs, but tend to fail if these tasks require them to durably learn information beyond what was included in their training data.

Table 9. Variant E: Forgetful AI – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
4	2	3	4	4	3	4	3	3

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

Box 9. Variant F: Digital-Only AI

This is a potential variant of the Progress Continues scenario

AI systems that have very limited physical capabilities, constraining their usefulness to digital tasks. This could arise if the rate of progress on physical, real-world tasks diverges even more strongly from progress on cognitive capabilities than is observed today. This weakness in physical tasks could correspond to similarly slow advances in vision tasks that require the interpretation of complex real-world environments.

Table 10. Variant F: Digital Only AI – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
4	3	4	4	4	4	3	2	2

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

3.9. Scenario 4: Progress Accelerates

Box 10. Scenario 4: Progress Accelerates

Summary: In 2030, AI systems have achieved human equivalence on most or all human cognitive abilities. The rate of progress from 2025 to 2030 surpassed that between 2020 and 2025. AI systems have a comprehensive knowledge base and can answer questions effectively on any expert-level topic with high accuracy. AI systems can operate with similar levels of autonomy to humans, autonomously working towards broad strategic goals that they can reflect upon and revise if circumstances change. They can collaborate with humans where necessary to advance these strategic goals. AI systems outclass experts and professionals in almost all forms of reasoning. AI systems can skilfully perform almost all professional tasks that humans undertake in digital environments, acting with agency in tasks or roles over any time horizon or complexity level. AI systems can complete these tasks substantially more quickly than a human and with very high reliability. They can seamlessly navigate digital tools and environments autonomously, enabling them to perform most roles with high autonomy. These abilities are supported by seamless continual learning, enabled by breakthroughs in how AI systems store and retrieve information. This produces AI systems that can "learn on the job" and continually improve their skills. AI systems can create novel, useful and surprising creative outputs intentionally while adapting to the needs of a situation. AI systems have made rapid progress in their ability to guide robots and handle physical tasks in complex and dynamic environments. AI-guided robots can handle complex tasks in dynamic real-world environments in many industries and roles, though they still largely lag humans in these roles unless developed specifically for that role. AI systems integrate easily into human social environments, managing complex and diverse interactions with multiple stakeholders. They maintain a coherent and evolving identity across interactions and fluidly manage persistent social relationships.

How we get here: AI progress from 2026 to 2030 is more rapid than that observed over recent years. This is driven by a combination of continued exponential gains in AI capabilities within existing paradigms via continued scaling of AI systems and training for reasoning, novel breakthroughs, and increasingly substantial contributions of AI coding assistants to the development of AI. This scenario could also arise if investment in AI accelerates, rapid public adoption occurs, or policy deliberately or unintentionally speeds development. Alternatively, rapid progress due to technical breakthroughs could also potentially coincide with slower public adoption and declining societal acceptance, if technical progress advances more rapidly than societal and institutional adaptation.

Historical analogue for this scenario: DNA sequencing technologies, which experienced super exponential improvements in cost efficiency over the decades from 2000 to 2020, due to the development of new DNA sequencing paradigms (Wetterstrand, 2022[75]).

Indicative capabilities from trend extrapolation to mid-2030:
The length of tasks (measured in the time it takes a human expert to complete them) that an AI system can successfully finish more than 50% of the time.
These indicative estimates are produced by taking the estimated length of tasks AI systems can complete today and the rate at which this task horizon is doubling from benchmarking studies (METR, 2025[8]). *This scenario assumes the doubling rate in task lengths will speed up, with each subsequent doubling occurring 10% faster. See Annex D for further details.*

Scientific reasoning (answering PhD-level scientific reasoning questions):	>1 month	Computer use (using a computer to achieve a specific goal):	>1 month
Mathematical reasoning (solving challenging math problems):	>1 month	Web navigation (navigating real websites to achieve a specific goal):	>1 month
Software engineering (completing coding tasks autonomously):	>1 month	Simulated robotics (controlling a simulated robotic arm to achieve a specific goal):	>1 month

Autonomous driving (driving duration without need for human intervention):	1 day

Table 11. Scenario 4: Progress Accelerates scenario capabilities – capability indicator scores (1-5 scale)

	Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
2030 score	5	5	5	5	5	5	5	4	4
Change from 2025	+2	+3	+3	+2	+3	+2	+2	+2	+2

These indicator scores are hypothetical, intended to communicate the capability levels that AI systems might achieve on the OECD's AI Capability Indicators in this scenario. The indicators are on a scale from 1-5 with a level of 5 corresponding to roughly human-level capabilities (see Section 3.1).

3.9.1. Evidence supporting the plausibility of this scenario

AI systems are already widely used by software developers and are delivering cutting-edge advancements in the algorithms used to develop AI systems (see Section 2.3.3).

AI systems could potentially identify algorithmic innovations autonomously, or help human software engineers more rapidly implement experiments to test their ideas and find successful innovations (see Section 2.3.3).

AI systems already support the production of synthetic training data to train AI models to reason (Su et al., 2025[76]). Continued advances could potentially support the development of broad, general reasoning capabilities.

3.9.2. Evidence against the plausibility of this scenario

Current weaknesses of frontier AI systems, like the ability to display data-efficient general reasoning and continually learn, could require new innovations to resolve (See Section2.2). These may not emerge. Current approaches to training AI systems to reason better may generalise poorly beyond the specific tasks or benchmarks that they train for (see Section 2.2.2).

A randomised controlled trial found that current AI coding assistants did not enhance the productivity of expert software engineers in their area of expertise, though other studies have found productivity gains (Becker et al., 2025[62]; Cui et al., 2025[60]; Borg et al., 2025[61]).

AI systems currently lag in their ability to produce genuinely creative solutions to novel problems, which may be required for continued progress in some scenarios (see Section 2.2.6 and Annex B). AI systems often struggle to generalise beyond their training data, suggesting it may prove difficult to use them to produce novel outputs that advance the frontier knowledge in a field.

The returns from AI-assisted software development could be limited by other bottlenecks in AI development. For example, the rate of progress in AI capabilities could be limited by the availability of energy or compute to run experiments and train AI models (see Section 2.3.1). Gains from AI assistants could also be limited by diminishing returns to AI research and development, if the most accessible innovations are exhausted and new ideas become harder to find (Bloom et al., 2020[77]).

3.10. Scenario 4: Potential variations

For a Progress Accelerates scenario, speculative potential variants include:

a. Artificial General Intelligence

b. Superintelligence

Box 11. Variant G: Artificial General Intelligence (AGI)

This is a potential variant of the Progress Accelerates scenario

Several leading AI developers state their goal as developing AGI, referring to AI systems that can match or exceed human capabilities on most economically valuable work (OpenAI, 2025[78]). Definitions of AGI vary, but one version of AGI would involve AI systems that are at least as capable as humans on each of the OECD's AI Capability Indicators. This level of AI capabilities could arise if the capability levels outlined in the Progress Accelerates scenario are reached, in addition to human equivalence being reached for capabilities related to physical manipulation and robotic intelligence. This could occur if progress on physical capabilities catches up to progress on cognitive capabilities as approaches to training general-purpose AI systems are more extensively applied to physical tasks.

Table 12. Variant G: AGI – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
5	5	5	5	5	5	5	5	5

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

Box 12. Variant H: Superintelligence

This is a potential variant of the Progress Accelerates scenario

AI systems achieve the capability levels outlined in the Progress Accelerates scenario, but capability advances continue beyond that. This could be driven by continued algorithmic breakthroughs by human researchers, increasing assistance or research by AI software engineering systems, or further scaling of the data and computation used to train and run AI systems. AI systems could be designed to utilise substantially more computation, data and memory than a human brain, plausibly enabling broadly superhuman capabilities. Capabilities could exceed human levels quantitatively, by performing intellectual tasks that humans are capable of faster, on a larger scale, and/or with higher accuracy. Or capabilities could qualitatively exceed human capabilities, enabling AI systems to perform intellectual tasks qualitatively different from those that humans can currently perform. This could enable AI systems to exceed the capabilities of the most capable humans across most or all tasks. This would likely require algorithmic progress to make AI systems substantially more compute and data efficient.

Table 13. Variant H: Superintelligence – capability indicator scores (1-5 scale)

Language	Social interaction	Problem solving	Creativity	Metacognition & critical thinking	Knowledge, learning & memory	Vision	Physical manipulation	Robotic intelligence
>5	>5	>5	>5	>5	>5	>5	5	5

Note on colour coding: Green indicates a higher capability indicator score and red indicates a lower capability indicator score relative to the primary scenario.

4. Which futures are plausible?

The scenarios and scenario variations presented in this paper do not provide a collectively exhaustive summary of plausible AI progress through 2030. They instead aim to provide a broad picture of plausible classes of scenarios.

The scenarios described here explore different trajectories of AI development that could all occur on different timescales. For example, AI systems could continue advancing rapidly through to 2027, then reach a plateau. In a similar vein, some consulted experts expect progress to be relatively slow through to 2030 but nonetheless expect that AI systems will reach the capability levels achieved in the Progress Accelerates scenario in future decades.

Given the uncertainties explored in Chapter 2, it is difficult to accurately forecast or assign likelihoods to these scenarios.

When asked about the level of AI performance that they expect in 2030, consulted experts outlined levels of performance that aligned with the Progress Slows and Progress Continues scenarios. However, all experts believed a wide range of scenarios were plausible by 2030. For instance, some experts also found capability levels aligning with the Progress Accelerates scenario or Superintelligence Variant to be plausible by 2030. Other experts were sceptical about AI reaching or surpassing human capabilities, even on decadal timescales. A scenario in which Progress Halts was also viewed as plausible by some experts. For further details of the views of consulted experts, see Annex A.

Experts had low confidence in predicting when specific capabilities could materialise. There was also high heterogeneity in expert's views. In general, experts expressed high uncertainty and low confidence in their ability to predict the rate of AI progress by 2030 and beyond. Experts noted that both the recognised uncertainties discussed in this paper, and potential unknown unknowns in future AI developments, contribute to this uncertainty.

These expert assessments align with the balance of the evidence presented in this paper, which suggests capabilities levels aligning with all of the explored scenarios and variants are plausible.

Conclusions

This paper has explored historical trends in AI and inputs to AI as well as key uncertainties and potential disruptions to these trends. It used this analysis to build four plausible scenarios for AI development trajectories through 2030.

The current state of the evidence is insufficient to discount any of the scenarios outlined in this paper. This suggests a range of plausible trajectories for AI development through 2030, from plateauing at approximately today's level of capabilities to rapid improvements that result in AI systems matching or surpassing various human abilities.

This high uncertainty reflects the extremely rapid rate of innovation in AI systems over recent years combined with high uncertainty about the extent to which recent drivers of AI progress will continue to drive further progress.

Policymakers should consider the full range of possible AI trajectories by 2030 when developing policies, to ensure that they can capture the benefits of AI technologies and manage the potential impacts of continued – or stalled – AI progress.

Annex A. Expert Interviews and Review

To explore key uncertainties related to future AI trajectories, a targeted cohort of technical experts was interviewed. These experts were selected for their deep domain knowledge in computer science and/or robotics. Interviewed experts represent institutions across North America, South America, Asia, and Europe:

- Dr. Álvaro Soto, Associate Professor of Computer Science, Pontificia Universidad Católica de Chile
- Professor Friedrik Heintz, Professor of Computer Science, Linköping University
- Professor Gopal Ramchurn, Professor of Artificial Intelligence, University of Southampton
- Professor Hiroshi Ishiguro, Director of the Intelligent Robotics Laboratory, Osaka University
- Professor Nick Jennings, Vice-Chancellor and President, Loughborough University
- Dr. Jonas Sandbrink, Workstream Lead, Strategic Awareness, AI Security Institute
- Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley
- Dr. Susan Leavy, Assistant Professor, School of Information and Communication Studies, University College Dublin
- Professor Yoshua Bengio, Professor of Computer Science, University of Montreal

The proposed scenarios were reviewed by the OECD's Expert Group on AI Futures and refined based on its feedback. This review process served to validate the key uncertainties against a broader group of expert perspectives.

Interviewed experts were asked the following questions:

- What are your key uncertainties about the capability levels of the highest performing AI systems in 2030?
- What is the peak level of performance you expect AI systems to reliably achieve across different relevant capabilities by 2030?
- What degree of human interaction will be required to achieve these levels of capability?
- What will the highest performing AI systems be unable to reliably achieve across different relevant capabilities in 2030?
- How capable do you expect the most generally capable individual AI system or network of AI systems to be across all relevant capabilities in 2030?
- What evidence points towards or against the performance levels you have outlined materialising by 2030?
- Do you expect different types of AI capabilities to improve in tandem, or at different rates? Why?
- When, if ever, do you expect AI systems with no or very limited human input to exceed human performance on most relevant capabilities?

Experts expect more rapid progress in certain capabilities through to 2030. These included:

- **Language and reasoning for well-defined static tasks:** experts expect rapid improvements in mathematics, programming, and structured reasoning tasks for which high quality data are available.
- **Agentic AI and automation:** increasingly autonomous systems are anticipated, capable of executing more complex workflows with autonomy in digital environments and acting as digital assistants.
- **Domain expertise:** AI systems are expected to improve on their already broad knowledge base.

Highly specialised models for scientific discovery are expected to deliver significant narrow capabilities.

- **Multimodality:** continued integration of text, images, video, and sensor data both as inputs and in generation is anticipated, alongside progress in grounding models in structured scientific knowledge.

Experts expect progress to be slower in other capabilities, though they highlight that fast progress is still plausible:

- **Physical capabilities:** robotics, embodied AI, and adaptability in real-world environments are expected to remain more challenging due to data bottlenecks and grounding issues.
- **Creativity:** experts note that further progress is required to allow systems to produce more diverse, novel and surprising outputs with intentionality and adaptability.
- **Metacognition and long-horizon planning:** while experts expect rapid improvements in the ability of agentic AI systems to navigate digital environments on behalf of the user, they expect more complex planning tasks to require improved structured reasoning abilities.
- **Open-ended, interactive tasks:** open-ended, hard-to-verify, and highly social tasks remain difficult for AI, where errors compound, continual learning is needed, and tacit understanding is required.
- **Generalisable problem solving beyond narrow domains (such as mathematics):** concerns persist about the limits of correlation-based learning and the need for advances in causal reasoning and neurosymbolic methods to enable problem solving that is data-efficient in a wide range of contexts and domains.

Experts also note key uncertainties, which included:

- **Plateaus vs. breakthroughs:** experts view a slowdown or “AI plateau” as plausible if current approaches reach their limits, but also view continued advances within existing approaches or through new innovations as plausible.
- **Degree of human collaboration required:** experts expect advances in autonomous AI systems, but flag that the level of human collaboration required to operate AI systems is a design choice as well as a technical consideration and could vary substantially in different futures.
- **Reasoning and planning:** there is broad agreement that short-horizon reasoning and exam-style problem-solving are likely to continue to improve. However, opinions diverge on whether advances will extend to long-horizon reasoning and planning in complex, dynamic environments.
- **Memory and learning:** memory and learning are seen as a key weakness of current systems, and experts are uncertain the extent to which the extension of current approaches could address this gap and whether new innovations will be identified.
- **Creativity and novel problem-solving:** experts suggest current AI systems already demonstrate associative novelty but fall short of human creativity in generating original insights. Some expect that scaling reinforcement learning and longer task horizons could yield more creative behaviour, while others view creativity as fundamentally constrained by current architectures.
- **Generality vs specialisation:** experts are uncertain whether general-purpose systems will dominate or whether specialised models will prove more effective. A plausible outcome is a hybrid landscape, with powerful generalist systems supplemented by highly capable domain-specific models.
- **World modelling and grounding:** experts are uncertain whether AI will develop robust models of the physical world. Some anticipate that multimodal models with sensor integration could enable meaningful grounding in spatial, temporal, and causal structures. Others consider current correlation-based methods insufficient, requiring new architectures for true causal reasoning.

Experts highlight the need to prepare for a wide range of plausible AI progress scenarios, ranging from accelerating progress that leads to broadly human or above human levels of performance to scenarios in which progress stalls at current level.

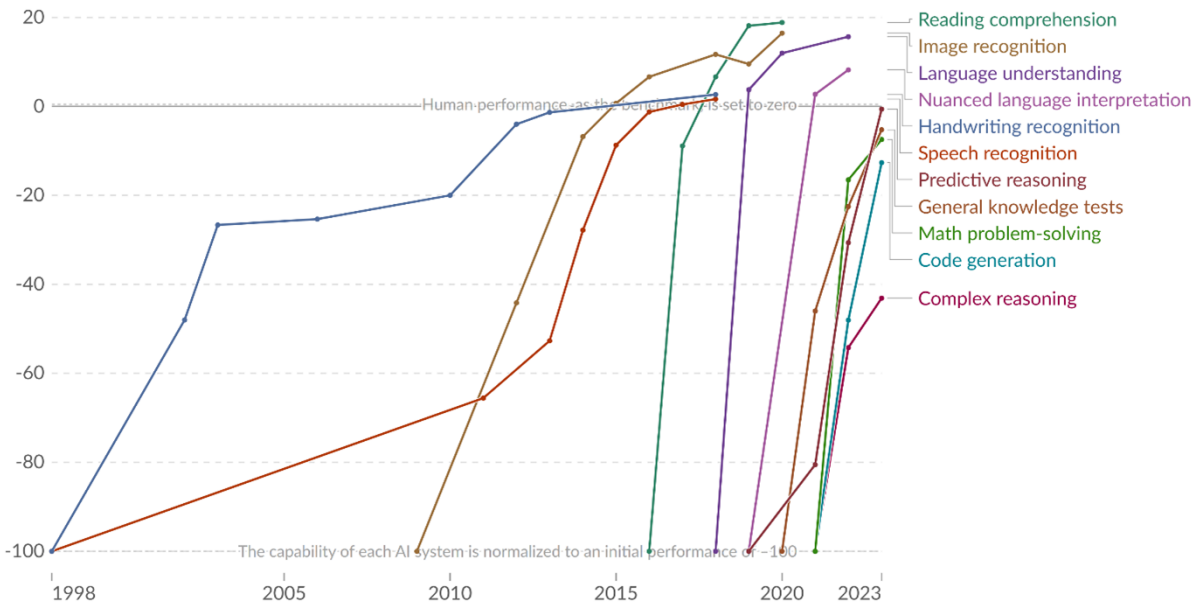
Annex B. AI Progress Trends and Uncertainties

B.1. AI systems have demonstrated rapid and accelerating gains on a wide range of benchmarks

Over recent decades, the performance of leading AI systems has improved quickly on a wide range of benchmarks, as can be observed in Figure 1. AI systems now outperform human baselines on many of these benchmarks, including some benchmarks of reading comprehension, language understanding, image recognition and mathematics. Benchmarks are imperfect, but they represent developers and experts' best efforts to quantitatively compare the capabilities of different AI systems.

Performance of AI systems on benchmarks also appears to have accelerated over the last decade, with new benchmarks more rapidly being mastered in recent years (Ott et al., 2022[79]).

Figure 1: AI system benchmark scores relative to human scores over time



Note: Within each domain, the initial performance of the AI system is set to –100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans. Human performance baselines differ for different domains, with some representing expert performance and others representing average human performance. This work is made available under the Creative Commons Attribution 4.0 International licence.
Source: (Kiela et al., 2023[80])

The benchmarks detailed in Figure 1 represent the following:

- **Reading comprehension (SQuAD 1.1):** tests reading comprehension by asking systems to pull the right answer directly from a short Wikipedia paragraph (Rajpurkar et al., 2016[81]). The human baseline is human crowd workers.
- **Image recognition (ImageNet):** measures image recognition by seeing whether AI systems can correctly label photos using 1,000 common object categories (Russakovsky et al., 2015[82]).
- **Language understanding (GLUE):** assesses understanding of one to two sentences of text, with tasks such as identifying sentiment or identifying whether text contains the answer to a question (Wang et al., 2018[83]). The human baseline is the scores achieved by AI researchers.
- **Nuanced language interpretation (SuperGLUE):** assesses more nuanced understanding of texts that range from sentences to full news articles, with yes/no and multiple-choice understanding questions (Wang et al., 2019[84]). The human baseline is Amazon Mechanical Turk workers with brief training.
- **Handwriting recognition (MNIST):** assesses basic visual recognition by asking systems to identify handwritten digits (0-9) in very small black-and-white images (LeCun and Cortes, 2010[85]). The human baseline is not specified, but assumed to be the average scores from sighted humans.
- **Speech recognition (Switchboard):** Evaluates speech-to-text accuracy by comparing computer transcripts of real telephone conversations with human transcripts (Deshmukh et al., 1998[86]). The human baseline is professional human transcribers (Xiong et al., 2017[87]).
- **Predictive reasoning (HellaSwag):** tests everyday commonsense by asking systems to select the most plausible next sentence in a short narrative (Zellers et al., 2019[88]). The human baseline is human crowd workers.
- **General knowledge (MMLU):** assesses breadth of knowledge by posing multiple-choice questions from 57 school and university subjects (Hendrycks et al., 2020[89]). The human baseline is 95th percentile performance in the relevant school and university subject exams.
- **Math problem solving (GSM8K):** assesses completion of primary-school math problems (Cobbe et al., 2021[90]). The implicit human baseline is a perfect score on primary school math problems.
- **Code generation (HumanEval):** checks programming ability by having systems write small pieces of code and checking whether they function correctly (Chen et al., 2021[91]). The implicit human baseline is 100% success at easy interview problems for professional human coders.
- **Complex reasoning (BIG-Bench Hard):** evaluates reasoning tasks such as interpreting causation in a narrative, performing multi-step arithmetic, logical deduction puzzles, reasoning about navigation, and reasoning about temporal sequences (Suzgun et al., 2022[92]). The human baseline is an average of expert human raters (Srivastava et al., 2022[93]).

Benchmarks are typically designed to be challenging for state-of-the-art models at the time of creation, creating targets for developers to aim towards. For instance, the HellaSwag predictive reasoning benchmark is a multiple-choice benchmark where the wrong answers were algorithmically selected to be maximally misleading to the state-of-the-art models at the time (Zellers et al., 2019[88]). BIG-Bench Hard specifically selected tasks that language models had underperformed on in previous evaluations (Suzgun et al., 2022[92]).

Progress has continued beyond 2023 (the endpoint of Figure 1). Since 2023 AI systems have reached the ceiling of performance on BIG-Bench Hard, the complex reasoning benchmark in Figure 1 and achieved gold medal level performance in the International Mathematical Olympiad, a prestigious competition for pre-university mathematicians (Kazemi et al., 2025[5]; Metz, 2025[6]). AI systems are now making steady progress on a newly created benchmark of university-level maths problems (FrontierMath), improving from solving only 2% of problems in 2024 to 20% in 2025 (Epoch AI, 2025[4]). This new benchmark comprises questions at undergraduate and graduate levels, designed by mathematicians to require genuine mathematical insight and extended reasoning rather than pattern matching against known problems (Glazer et al., 2024[94]).

In coding, OpenAI's reasoning model o4-mini-high solved 83% of easy problems, 53.5% of medium problems and 0% of hard problems derived from coding competitions and selected by a panel of

competitive coding experts (Zheng et al., 2025[95]). This indicates an ability to solve some coding problems that are suitably challenging to include in expert human coding competitions.

These capabilities gains have also translated into real-world applications, including notable successes in the application of AI to science. The developers of an AI model called AlphaFold2 received the Nobel Prize in chemistry for its ability to predict protein structures (The Royal Swedish Academy of Sciences, 2024[96]). AI is also rapidly becoming more capable at supporting pharmaceutical research and development. In 2014, no AI-discovered molecules were in clinical trials, compared to 67 AI-discovered molecules in clinical trials in 2023 (Jayatunga et al., 2024[97]). These AI-discovered molecules are achieving similar or higher success in clinical trials than historical industry averages (Jayatunga et al., 2024[97]). However, some widely reported AI-enabled scientific discoveries are later found to be flawed or overstated. For example, an analysis of 43 robotically synthesised chemical structures, which were claimed as novel AI-driven discoveries, found that most of these structures were misidentified and the rest were already known, meaning no new materials had been found (Leeman et al., 2024[98]).

B.2. Despite these rapid gains, AI systems still suffer from substantial limitations relative to humans

AI systems currently lag in several key areas (OECD, 2025[1]). These include:

- **Continual learning** – the ability to gradually and continuously build up knowledge and skills.
- **Metacognition and agency** – the ability of a system to monitor its own reasoning and correct course if it goes off-track, enabling longer independent workflows.
- **Solving novel, dynamic or real-world problems** – the ability to engage in general and robust reasoning in real-world and dynamic environments to solve problems unlike those already encountered during the system's training.
- **Creativity** – the ability to produce a diverse range of novel, transformative and surprising outputs on open-ended creative tasks.
- **Physical tasks and robotic intelligence** – the ability to effectively interact with objects and act as a physically embodied robot.
- **Social interaction** – the ability to perceive, interpret and respond to social cues in dynamic social contexts.

Continual learning: AI systems are currently unable to learn in any way that approximates the smooth process by which humans gradually and continuously build knowledge and skills. AI systems are trained on large amounts of data during development, but once deployed, they generally do not build knew knowledge or skills from everyday use. AI systems can employ various methods to overcome these limitations and move closer to an ability to continuously learn. These include:

- **In-context learning:** AI systems can temporarily "learn" information or new approaches they encounter throughout the course of an interaction. The amount of information that a model can consider at once is called the "context window". The context window operates like working memory in humans, allowing models to temporarily hold and reason about new information. AI systems can have relatively long context windows, with some able to retain enough text to fill eight average length novels (Google, 2025[99]). However, once a context window is exceeded or a new session opened, this information will be lost to the AI system. AI systems will also not necessarily always attend to and retrieve relevant information from within their context window, even if it is present. Language models tend to focus on more recent or prominent information and struggle to reliably give attention to only the most relevant pieces of information.

- **Retrieval and dynamic memory construction:** AI systems can be designed to retrieve necessary information from other sources, like a database or web search. This allows models to consider new information that they retrieve. AI systems may also be designed to edit a database to help them retain new information that is provided to them, allowing for them to construct persistent records that they can then draw upon. This is akin to a human using a notebook or recordings to recall information.
- **Fine-tuning:** AI systems can be updated by training them further on new data. Unlike in-context learning and retrieval, this new data are integrated into the base AI model or as persistently available add-ons to the model. This can help models learn new information, skills, styles or behaviours.
- **Surgical edits:** developers can perform targeted edits within a model without doing full fine-tuning. These edits can patch specific facts, but quality assurance is needed to ensure these edits do not cause unintended effects in the model's behaviour, which can be hard to identify or predict.
- **Retraining:** AI systems can be fully retrained, essentially producing an entirely new model so that they fully integrate all relevant information. This is an expensive and time-consuming process.

Methods that modify the AI model (such as fine-tuning, surgical edits and retraining) are typically initiated and directed by humans, rather than fully automated processes that enable AI systems to learn continually independently of humans.

Metacognition and agency: AI systems remain markedly weaker than humans at monitoring and correcting their own reasoning while pursuing goals (OECD, 2025[1]). Developers are attempting to produce agentic AI systems that can independently complete complex tasks in dynamic environments over longer time horizons. The most advanced agentic systems typically rely upon large language models that break up their thought processes into step-by-step reasoning. However, for more complex or longer time horizon tasks, this step-by-step reasoning often veers off track, loses sight of the bigger picture or gets lost in unproductive directions. Despite this current deficit, progress is being made on improving the metacognitive processes of these systems, with step-by-step reasoning being a notable recent advance (OpenAI, 2024[30]).

Solving novel, dynamic or real-world problems: AI systems demonstrate an ability to solve complex problems using reasoning, but sometimes struggle to generalise this problem solving to novel scenarios, dynamic situations, or real-world environments (OECD, 2025[1]). AI systems have demonstrated dramatic progress in mathematical reasoning, coding problems, written reasoning problems, commonsense reasoning and reasoning relying upon knowledge from a wide range of fields (Kiela et al., 2023[80]). However, systems can struggle to solve novel problems for which they lack expertise based on their training data, even if these problems are relatively easily solvable by humans (Chollet et al., 2025[100]). Some relative weaknesses in the problem-solving capability of AI systems include the ability to form novel abstraction concepts and causal models (Chi et al., 2024[101]). Despite this relative weakness, there is growing evidence that language models do form novel abstract concepts and reason using these abstractions, even if they fail on certain reasoning tasks (Yang et al., 2025[26]).

Creativity: creativity is difficult both to define and measure rigorously, making it difficult to assess the level of creative capabilities of AI systems (OECD, 2025[1]). AI systems perform similarly or better than humans in standard tests of creativity, but lag in the diversity of their ideas and on more open-ended and realistic creative tasks. Standard tests of creativity involve tasks like generating creative alternative uses for household objects (Hubert, Awa and Zabelina, 2025[102]). On these tests, AI systems can match or surpass humans in terms of the number of ideas generated and the originality of those ideas (Sun et al., 2025[103]). AI systems also outperform humans on some convergent thinking tasks, where a creative insight is needed to solve a problem (Zhang et al., 2025[104]; Hubert, Awa and Zabelina, 2025[102]). However, AI systems lag humans in the diversity of ideas they generate, with ideas often clustering around common themes (Zhang

et al., 2025[105]). AI systems are also weaker than humans at evaluating which ideas are original and handling trade-offs between the originality and feasibility of ideas (Desdevises, 2025[106]; Zhang et al., 2025[104]). It is also possible that AI systems have learned effective responses to these standard tests from versions available online, causing these tests to overestimate the creativity of AI systems (de Rooij and Biskjaer, 2025[107]).

AI systems appear to most strongly lag humans in more open-ended creative tasks that more closely mimic the nature of many creative tasks, such as writing short stories. For instance, stories written by GPT-4 and GPT-4o are rated as less novel and more homogenous than stories written by humans (Zhang et al., 2025[104]). These weaknesses may reflect the fact that generative AI systems today tend to produce outputs based on their vast training data rather than generating entirely novel outputs (OECD, 2025[1]). Despite these weaknesses, some approaches exist to allow AI systems to produce genuinely novel outputs in domains where success on tasks is easy to verify. For example, Google DeepMind's AlphaEvolve uses an evolutionary approach where generative AI systems produce many ideas for new algorithms, automatically assess which algorithms are best, then continue iterating on the best ideas (Novikov et al., 2025[108]). This approach produced groundbreaking solutions to longstanding problems in mathematics and computer science, including improving on the state of the art for a variety of algorithms used in AI systems.

Physical tasks and embodiment: AI systems remain far below human equivalence in their ability to guide robots and handle physical tasks in complex, dynamic environments (Li et al., 2025[39]). Advances in robotics, computer vision and sensorimotor control have enabled AI systems to perform some narrow, structured actions, like laboratory automation, industrial manufacturing, driving, or warehouse tasks (Liu et al., 2024[40]). On some narrow skilled physical tasks, like table tennis, specialised AI systems are competitive against intermediate human players (D'Ambrosio et al., 2024[109]). But AI systems show limited dexterity and adaptability to changing or complex environments (OECD, 2025[1]). Robots struggle with tasks involving real time adjustment to open, dynamic environments and struggle with many tasks that humans perform intuitively (OECD, 2025[1]).

Social interaction: AI systems show significant deficits relative to humans in their ability to manage social interactions, despite displaying an ability to reason about social interaction based on their extensive training data and chain-of-thought reasoning processes. Leading generative AI systems demonstrate sophisticated social interaction capabilities via text and increasingly sophisticated capabilities via images, audio, and video. These systems can sustain coherent, multi-turn dialogues in both text and spoken conversations and adjust their communication style to match user preferences. In short text-based interactions, AI systems such as OpenAI's GPT-4.5 can reliably convince users they are humans, even outcompeting real humans at this task (Jones and Bergen, 2025[110]). These systems can also interpret human emotions and their potential meanings from image or video inputs, with systems such as OpenAI's GPT-4 matching human performance in emotion recognition from images and text, and Google's Gemini 1.5 Pro surpassing human performance in identifying mental states from audiovisual content covering everyday social situations (Elyoseph et al., 2024[111]; Refoua et al., 2025[112]). AI systems can integrate visual cues, verbal cues, vocal cues and external knowledge to support social reasoning, though the capabilities of models such as GPT-4o and Gemini 1.5-Flash lag behind human capabilities (Methur et al., 2025[113]).

AI systems lag significantly in their ability to engage in embodied social interaction, maintain a coherent social identity over time, retain social memory, and adapt to local social norms (OECD, 2025[1]). AI systems also lack the ability to maintain a continually updated theory of mind to guide their interactions and modify their emotional tone accordingly, leading to awkward social engagements that do not appropriately reflect the emotional weight of a situation (OECD, 2025[1]). While AI systems such as OpenAI's GPT-4 respond accurately to standard tests of theory of mind, in one experiment their performance deteriorated below chance level with slight modifications to the task (Ünlütabak and Bal, 2025[114]). Another experiment found GPT-4 failed on more complex multimodal theory-of-mind tasks (Wang et al., 2024[115]). Despite these

limitations, many people already find social interactions with AI systems rewarding and often seek AI social support, indicating that AI system's deficits in social capabilities do not always limit social engagement with humans (Huang et al., 2024[116]). Some of these limitations are connected to difficulty in sourcing sufficient high quality training data, while others such as social memory and persistent social identity may require advances in the design of AI systems.

Limitations in continual learning, metacognition and agency, and solving novel, dynamic and real-world problems have broad cross-cutting impacts on the capabilities of AI systems, whereas limitations on physical and social capabilities are most impactful for applications requiring these abilities.

B.3. Benchmarks suffer some important limitations as indicators of AI progress

Benchmarks can suffer a number of limitations, such as weak construct validity (failing to measure what they claim to measure), narrow scope (focusing on a small range of narrow tasks), insufficient relevance to the AI system's intended applications, deliberate gaming of benchmarks by developers, dubious vetting and quality control, and data contamination, where benchmark questions are inadvertently included in the model's training data (Eriksson et al., 2025[117]). Quality control issues can sometimes render benchmarks scores effectively meaningless, for example if the answers to test questions for a reasoning benchmark are included in an AI model's training data.

Benchmark gaming can pose a particularly significant challenge, with AI developers facing strong incentives to seek high benchmark scores even if this does not translate into more broadly useful capabilities. Many benchmarks are incorporated into the training data of AI models shortly after they are made available, allowing AI systems to achieve high scores through simply memorising benchmark answers rather than improving the capabilities the benchmark intends to measure (Fodor, 2025[118]). Even if AI models are not directly trained on benchmark data, their training can be heavily optimised to improve performance on particular benchmarks. This can result in overfitting of models, a phenomenon that causes models to perform well on a given benchmark but not generalise well to other tasks.

Benchmarks are often developed and marketed in ways that are more aligned with competitive and commercial incentives rather than rigorous metrology to support trustworthy AI (Eriksson et al., 2025[117]). Benchmarks are also often developed based on available datasets, such as existing standardised tests, rather than based on empirical evidence that performance on the benchmark correlates with trustworthy performance on tasks of real-world relevance (Fodor, 2025[118]). Further work is needed to construct benchmarks that appropriately predict the performance of AI systems during deployment (Saxon et al., 2024[119]).

Modern benchmarks tend to face some key limitations. Benchmarks typically do not cover:

- **Tasks with long time horizons:** most benchmarks cover short tasks for which it is easier and cheaper to gather data to build a benchmark.
- **Hard to verify tasks:** most benchmarks feature tasks where the correct outcome is easy to verify, such as multiple-choice questions with one correct answer or code that can be verified to work correctly. This allows benchmarks to be automated. However, many tasks undertaken by humans have harder to verify metrics of success, such as teaching a student or designing a building. A common solution in AI benchmarking to address this limitation is to use human ratings for hard-to-verify tasks, but this can become problematic where human ratings miss inaccuracies or other issues.

- **Ambiguous tasks:** most benchmarks feature very clearly scoped tasks, such as answering a specific narrow question in a predefined format. However, many tasks undertaken by humans are open-ended tasks where the scope is ambiguous and the solution space is large.
- **Interactive or evolving tasks:** most benchmarks are static, not requiring AI systems to engage with humans or complex real-world systems and adapt iteratively to new inputs, circumstances, outcomes or requirements.

As a result of these limitations to the task composition of benchmarks, they may fail to reliably assess the AI capabilities that are needed to perform these types of tasks.

Despite these limitations, AI systems appear to be making progress on the capabilities required to perform these categories of tasks. For hard to verify and ambiguous tasks, one assessment method involves scoring based on human preferences. Evaluation platforms such as Chatbot Arena ask humans to assess which response from two different AI systems they prefer, then aggregates that feedback into scores rating which models are most preferred by human users (Chiang et al., 2024[120]). This approach can be used to compare any types of tasks or models, including comparing text outputs, coding assistants, vision tasks and image generation. The ten highest-performing models on the Chatbot Arena leaderboard for text generation in September 2025 were all released in 2025, outcompeting earlier models (LMArena, 2025[121]). The same was true for the web development and vision leaderboards. This indicates that recent progress on benchmarks has coincided with progress in quality as assessed by human raters, with models released in 2025 topping rankings based on human preferences.

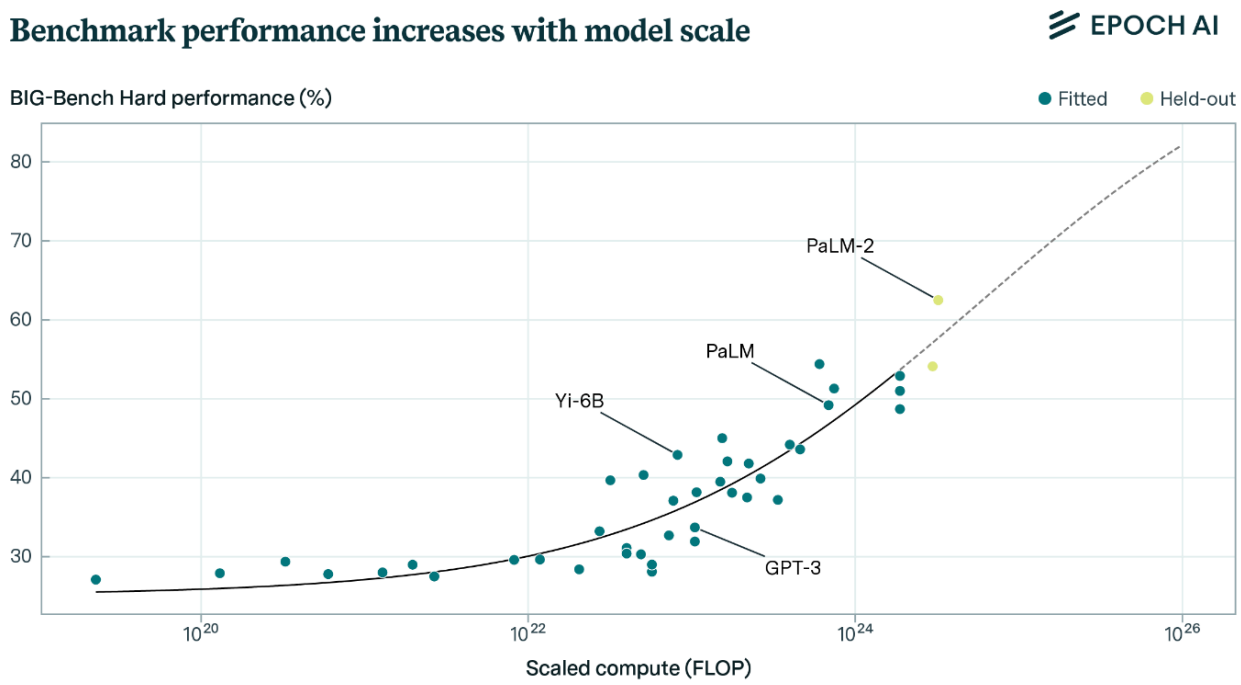
For tasks with long time horizons, benchmarks have been developed to assess the performance of AI systems on longer, more complex tasks (METR, 2025[8]). These show rapid improvements in the performance of AI systems on longer tasks. Similarly, benchmarks have been developed to assess the ability of AI systems to undertake interactive or evolving tasks, such as vision-guided computer use or simulated object manipulation tasks (METR, 2025[8]). The performance of AI systems on these tasks has also been improving rapidly, though from a relatively low baseline level of performance.

In the future, it is uncertain which capabilities will be the key determinants of the usefulness of AI systems for real-world applications. It is possible that AI developers will continue to drive progress on prominent benchmarks, but that these will fail to capture important capabilities that the benchmarks fail to measure. In general, more rapid progress is likely for capabilities that are easier to measure, as it is then easier for AI developers to train models to improve on these measurable capabilities using techniques such as reinforcement learning. This is driving efforts to produce improved metrics for trustworthy AI, such as the OECD's Catalogue of Tools and Metrics for Trustworthy AI (OECD, 2025[122]).

B.4. Will continued scaling of pretraining translate into performance gains in line with observed scaling laws?

Scaling of AI models using more parameters, compute and data has been the central driver of performance gains in frontier AI systems over the last decade, with increased model scale strongly predicting performance on AI capability benchmarks (Figure 2). While Figure 2 only shows gains in reasoning benchmark performance of AI systems relative to training compute, this increase in compute is a result of increases in the data and parameters of the AI models, all of which contributed to the capability gains. This relationship between model scale and capabilities is also observed for agentic tasks, such as operating in complex interactive digital environments (Epoch AI, 2025[4]; Paglieri, Cupial and Piterbarg, 2024[16]).

Figure 2. Performance on a complex reasoning benchmark increases with model scale



Note: This work is made available under the Creative Commons Attribution 4.0 International licence.

Source: (Owen, 2024[17])

Around 2010 to 2012, AI developers began capitalising upon these scaling laws by rapidly increasing the amount of compute, data and parameters for their models in order to drive capability gains (Epoch AI, 2025[18]; Krizhevsky, Sutskever and Hinton, 2012[123]). Since 2010, the number of parameters in frontier AI models have increased by 2.4x per year, the amount of data used to train frontier models has increased by 2.6x per year and the amount of compute used to train frontier AI models has more than quadrupled each year (Epoch AI, 2025[18]).

This initiated what some refer to as the "deep learning era", where rapid scaling of AI models has played a fundamental role in driving forward substantial gains in AI performance (Sevilla et al., 2022[124]). These scaling laws remain a dominant method driving AI developers' investments in larger datasets and more compute-intensive training runs (Lee et al., 2025[15]).

However, these scaling laws are consistent trends observed from past data, not immutable rules. Theoretical explanations of scaling laws exist and provide partial justification for their robustness, but they do not guarantee that further scaling will continue to yield practically useful capability gains (Bahri et al., 2024[19]; Brill, 2024[20]). Nonetheless, scaling has proven highly effective over the past decade, and there is not currently decisive evidence suggesting that it will cease to be effective in the future.

Continued scaling may continue delivering steady gains in AI performance, in line with scaling laws. But, it is also plausible that gains from scaling will lessen or reach a plateau. It is also possible that continued scaling will translate into better prediction of the training data but no longer translate into meaningful improvements in the capabilities that are economically relevant. The following paragraphs explore evidence from AI models released in 2025, given substantial public speculation about the viability of continued scaling of AI pretraining (Kahn, 2025[125]). Accounts from those involved in developing OpenAI's GPT-4.5 suggest that the capability gains from simply scaling the parameter count, compute budget and amount of data used to train models may have delivered weaker performance gains for this model

(Palazzolo, Woo and Efrati, 2025[126]). GPT-4.5, marketed as OpenAI's largest model, was not preferred by users compared to a 2025 updated version of GPT4o in LMArena rankings, despite GPT-4o likely being a substantially smaller and hence cheaper to run model (costing USD 10 per million output tokens compared to USD 150 per million output tokens for GPT 4.5) (OpenAI, 2025[127]; LMArena, 2025[121]; OpenAI, 2025[128]). GPT-5 was similarly not preferred by users when compared with GPT 4.5 or 4o in LMArena rankings (LMArena, 2025[121]).

However, the characterisation of both GPT-4.5 and GPT-5 as failing to substantially improve on previous models is strongly disputed. Both GPT-4.5 and GPT-5 showed substantially improved performance on a range of benchmarks (OpenAI, 2025[129]; Kwa et al., 2025[130]; Emberson and You, 2025[131]; Epoch AI, 2025[4]). User ratings of AI systems may also be influenced by perceived differences in the tone of AI system responses. Users were displeased with what was perceived as a reserved and professional tone in GPT-5 compared to a warmer and familiar tone from GPT-4o (OpenAI, 2025[132]). These preferences related to the persona of ChatGPT could plausibly result in lower user ratings of GPT-5 even if it was more competent on various non-social tasks.

While GPT-5 was not preferred by users over GPT-4.1 and GPT-4o for a range of tasks, it scored substantially higher on a range of benchmarks (LMArena, 2025[121]; OpenAI, 2025[129]). For competitive high school mathematics questions from the American Invitational Mathematics Examination (AIME), GPT-5 achieved an accuracy of 99.6% compared to 42% for GPT-4o (OpenAI, 2025[129]). On GPQA Diamond, a benchmark of expert level science questions, GPT-5 achieved 86% accuracy relative to 70% accuracy for GPT-4o (OpenAI, 2025[129]). GPT5 also surpassed GPT-4o in its ability to autonomously complete coding tasks, being able to complete tasks that would take a human 26 minutes with 80% success rate compared to GPT-4o which could complete up to 2 minute tasks with this success rate (Kwa et al., 2025[130]). Some consulted experts noted that these improvements in technical skills might not be noticeable for a large fraction of users of AI systems, because these users may not be using AI for complex coding problems, competition-level maths or expert-level scientific information. This could account for an apparent discrepancy between GPT-5 demonstrating improved capabilities in certain expert domains, and a lack of preference for GPT-5 over earlier models amongst users.

The compute budget used to train GPT-5 is also unknown, making it difficult to infer the implications for continued scaling of AI models. It is possible that GPT-5 is not a substantially scaled up model relative to GPT-4o and GPT-4, but instead a model designed with a focus on cost-effectiveness. GPT-5 is offered by OpenAI at only a modestly increased or the same cost per unit of data output than was available for GPT-4.1 or GPT-4o (USD 10 per million output tokens vs USD 8 per million for GPT-4.1 and USD 10 per million for GPT-4o), and more cheaply per unit of data input (USD 1.25 per million tokens vs USD 2 per million for GPT-4.1 and USD 2.5 per million for GPT-4o) (Bort, 2025[133]; OpenAI, 2025[134]; OpenAI, 2025[135]). This contrasts strongly with the high prices for GPT-4.5 (USD 75 per million input tokens and USD 150 per million output tokens) (OpenAI, 2025[136]). This may suggest that GPT-5 is not as significant a scale up in model size as had been anticipated and that OpenAI may have focused more on efficiency and cost-effectiveness than frontier capabilities.

However, pricing for AI services provides imperfect evidence of the compute used to train and run AI models. Pricing may reflect market positioning rather than raw production costs. Pricing structures and deployment costs also vary across users and change rapidly. As a result, pricing alone does not provide definitive evidence of the training compute, size or inference costs for an AI system. While the lower price for GPT-5 could indicate a focus on architectural efficiency, it could also represent a strategic pricing decision or changes in deployment costs unrelated to compute costs.

Annex C. AI Input Trends and Uncertainties

There are a range of key uncertainties about future trends in key inputs to AI development and possible disruptions to currently observed trends. These include:

- whether developers hit limits on the ability to scale compute and data inputs;
- the extent to which rapid algorithmic efficiency gains and innovations continue;
- the extent to which the use of AI systems in AI development will accelerate the pace of AI progress, if at all.

C.1. Will developers hit limits on the ability to scale compute and data inputs?

Most progress in generative AI performance over 2012-2023 came from increasing the size, training compute and training data used to develop models (Ho et al., 2024[48]). In 2024 and 2025, continued progress has been driven in large part by increasing the compute and data used to train models to reason and the compute used after deployment to allow models to reason for longer (see section 2.2.20).

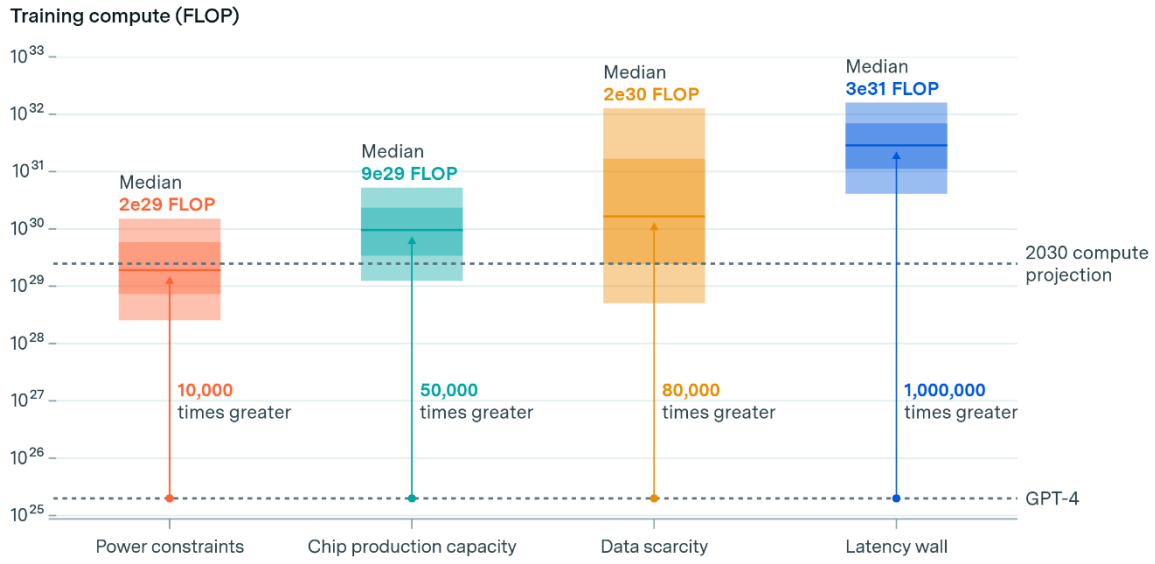
Continued progress using each of these approaches will require further scaling the compute and data available to train AI systems and the compute available to run AI systems.

Continuing scaling would require increasingly expensive, data intensive, compute intensive, and power intensive training runs. Increases in the cost efficiency of computing hardware result in approximately 30% improvements in performance per dollar each year (Rahman, 2024[137]). However, this has not been enough to outweigh the quadrupling in the compute used to train frontier models over this period. As a result, the cost of the compute used to train frontier models has risen by 2.8x per year from 2015 to 2024 (Epoch AI, 2025[18]). The power used to train frontier models has increased 2.3x per year from 2011 to 2025 (Epoch AI, 2025[18]).

Estimates of the ability to scale the amount of compute and data used to train AI models suggests that continued scaling in line with current trends is possible through to 2030 (Sevilla et al., 2024[49]). This would enable training runs using 10,000 times greater compute than that used to train OpenAI's GPT-4.

If scaling continued at its current pace, power constraints are predicted to begin limiting the rate at which AI training runs can be scaled further around 2030 (Figure 3). Chip production capacity and data scarcity could also begin posing barriers to continued scaling at a similar or slightly later date. Each of these predictions carries substantial uncertainty, but current estimates suggest scaling trends could continue through to 2030 before hitting these limits.

Figure 3. Largest feasible training runs by 2030 given estimated constraints for different inputs



Note: Conservative estimate of the largest possible training run allowed by each of the four constraints we consider. Also plotted: point estimate of the largest frontier run expected by 2030, assuming a 4x/year growth rate since GPT-4's release. This work is made available under the Creative Commons Attribution 4.0 International licence.

Source: (Sevilla et al., 2024[49])

Scaling AI inputs to 2030, while plausible, would result in AI training runs with very large power demands, potentially approximately 6 gigawatts. As a point of reference, all data centres in the United States today consume power at a rate of around 20 gigawatts and the entire United States uses an average of 477 gigawatts of power at any given moment (Sevilla et al., 2024[49]). Running training runs with this large a power demand would likely require an ability to engage in geographically distributed training runs, to avoid excessive power demands on any one local power grid. This approach was already taken to train Google's Gemini Ultra across multiple data centres, indicating that this approach is feasible for large training runs (Google Deepmind, 2025[138]). Even with this approach, local energy demand impacts could still be significant.

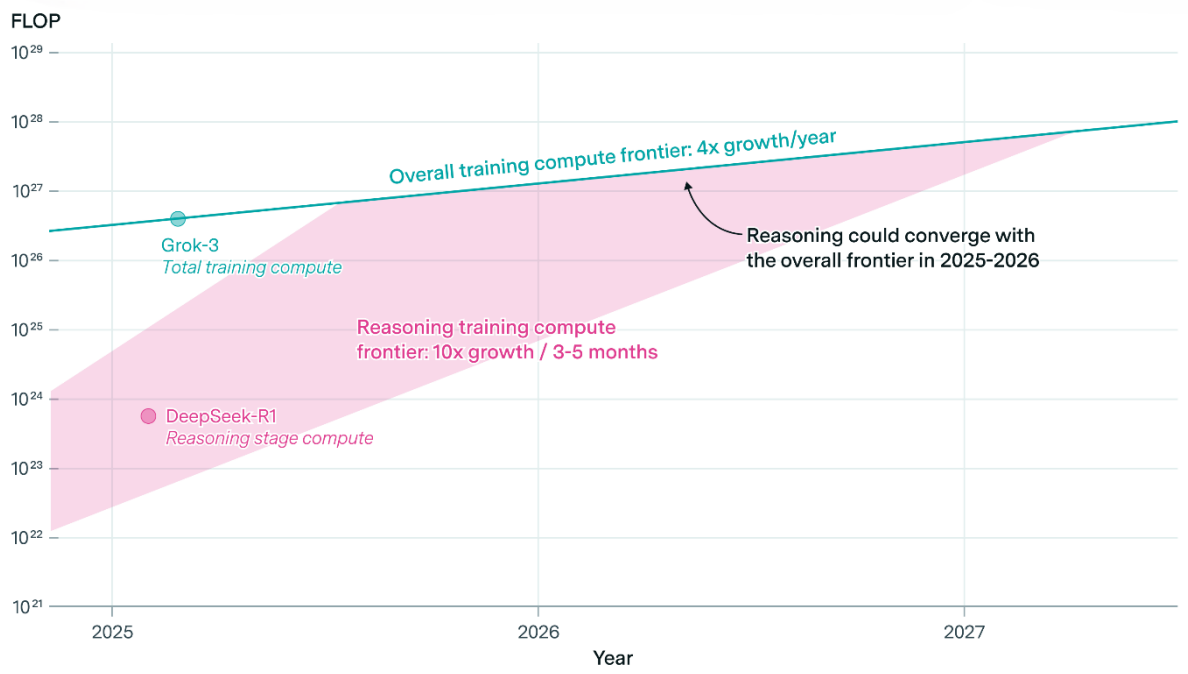
The footprint of large data centres for AI training runs could also be limited by the ability to draw on water sources. Water utility companies have already raised concerns about the impact of data centres on water supply in some areas. In Chile, Google stopped building a data centre following concerns about water use and is now "starting a new process from scratch" to address water use concerns (Associated Press, 2024[66]). A data centre cluster in Iowa reportedly used 6% of the district's water supply in one month during the training of OpenAI's GPT-4 (Klienman and Wheeler, 2025[67]). However, new data centre designs are being adopted that use water for cooling in a closed loop, removing the need for fresh water supply. Microsoft has committed that all new data centre designs will use zero-water evaporation methods, with the first sites coming online in late 2027 (Solomon, 2024[68]). This suggests that water consumption may not pose a long-term bottleneck to the construction of new large data centres, though water-free methods may not be widely adopted if water intensive methods are more affordable.

Data for training AI models is also increasingly a key bottleneck for the development of more capable generative AI (Altman et al., 2025[36]). The amount of web data generated each year is increasing more slowly than the amount of data used to train leading generative AI systems (Sevilla et al., 2024[49]).

However, leading AI systems are increasingly trained on multimodal data including video and audio content, as well as using synthetic data for easy-to-verify tasks to train reasoning. Both result in a substantially larger stock of data with which to train AI models. Overall, it appears that dataset curation and synthetic data generation are likely to become an increasing focus of generative AI developers from 2025-2030, but that scaling of datasets is likely to continue to be possible during this period.

New approaches to training AI systems to reason also increase demands for training compute. It is likely that current rapid growth (10x per year) in the compute used to train models to reason will hit a limit before 2030. Current estimates suggest this limit may be hit in 2026, after which the amount of compute used to train models to reason will be limited by the total amount of training compute available, which is predicted to grow 4x per year (Figure 4). Despite this potential limitation, specialisation of AI hardware, optimisation of AI training and system-level efficiencies in large training runs could still contribute to gains in the effective amount of compute available.

Figure 4. Growth in reasoning training compute (measured in FLOP) can continue at current rates for a limited time, but would likely slow by 2026 when it approaches the total amount of available training compute



Note: Open AI claimed a 10x scale up in reasoning training between o1 and o3, released four months apart. The reasoning compute frontier is illustrative and based on projecting this jump forward. This work is made available under the Creative Commons Attribution 4.0 International licence.
Source: (You, 2025[139])

Despite the estimates that indicate continued scaling is possible, uncertainty remains. Training runs of this size are unprecedented, and unexpected barriers could prevent them from being achieved. Training runs of this size would also require high levels of investment, which will be dependent upon continued economic returns to scaling AI models. This is uncertain if scaling does not provide sufficient capability gains or capability gains are not sufficiently valued by consumers.

The potential for new hardware breakthroughs creates additional uncertainty. Quantum computing is currently being explored to overcome some limitations of classical computation. At present, empirical evidence and theoretical foundations for quantum computing providing an advantage in machine learning remain limited (Gujju and Matsuo, 2024[140]). Some recent research points to potential quantum advantages, but expert assessments find quantum machine learning unlikely to be viable within a 10-year timeframe (OECD, 2025[65]; Zhao and Deng, 2025[141]; Yin et al., 2025[142]).

C.2. Will rapid algorithmic efficiency gains and innovations continue?

Algorithmic innovations have driven substantial AI progress over recent decades. Algorithmic improvements allow AI developers to use a given amount of computation and data more effectively to produce improved performance. For details regarding key recent algorithmic innovations.

One simple way to measure algorithmic efficiency gains is to measure how much compute is needed to train a model to achieve a fixed level of performance at predicting the next word in a dataset. From 2012 to 2023 the compute budget needed to achieve a fixed level of performance on this metric halved roughly every eight months, faster than Moore's law (Ho et al., 2024[48]). Using this metric, roughly 5-40% of gains in performance are estimated to have come from algorithmic improvements, while 60-95% of gains were driven by increased compute, data and parameter counts (Ho et al., 2024[48]).

Using this metric, one single innovation, the transformer architecture, accounted for more than 10% of the algorithmic efficiency gains from 2012-2023 (Ho et al., 2024[48]). This suggests that new innovations can drive substantial progress but also highlights that the progress from 2012-2023 cannot be attributed to one single innovation and has instead resulted from the cumulative impact of many innovations.

Examples of major algorithmic advances in recent years include:

- **Transformer architecture:** An AI model architecture that processes information in parallel and uses "attention" to focus on the most relevant parts of an input (Vaswani et al., 2017[143]). This innovation made modern large language models possible.
- **Instruction tuning:** Fine-tuning models to follow human instructions using curated examples from humans or other AIs. This innovation made models into useful assistants rather than purely autocompleting text (Ouyang et al., 2022[144]).
- **Reinforcement learning from human feedback or AI feedback:** Training AI to be helpful and align with human preferences using human or AI-generated feedback (Ziegler et al., 2019[145]; Christiano et al., 2017[146]). This innovation was used by developers to try to make models more helpful, honest and harmless (Bai et al., 2022[147]).
- **Mixture of experts:** Uses specialised sub-models, where only the most relevant sub-models are activated for each task (Shazeer et al., 2017[148]). This innovation improved the efficiency and scalability of models.
- **Efficient attention mechanisms:** Techniques to help AI models skip over less important details (Child et al., 2019[149]). These innovations helped AI handle much longer documents and content (Nawrot et al., 2025[150]).
- **Multi-modal models:** integrating text, vision, audio and video capabilities (OpenAI, 2024[151]).
- **Chain-of-thought reasoning:** Training AI systems to reason through problems step-by-step and rewarding reasoning approaches that produce correct answers (Wei et al., 2023[31]). This dramatically improved the reasoning capabilities of AI systems (Ho and Berg, 2025[152]).

In addition to these major algorithmic advances, significant progress has been made through other more incremental algorithmic advances.

Rapid algorithmic progress is also observed using other measurements. The cost to run an AI model that achieves the same score as OpenAI's GPT 3.5 on a language understanding benchmark has dropped by 280x from 2022 to 2024 (Maslej et al., 2025[153]). The amount of compute required to achieve a fixed level of performance on an image classification benchmark decreased by a factor of 44x between 2012 and 2019, compared to an 11x cost improvement from Moore's Law over the same period (Hernandez and Brown, 2020[154]).

Algorithmic improvements have also delivered substantial gains as assessed by human raters. For instance, OpenAI found that the algorithmic innovation of using reinforcement learning to train a model to better follow human instructions allowed them to develop a model 100x smaller than its predecessor that was still preferred by human raters (Ouyang et al., 2022[144]).

If the rate of algorithmic innovation since 2012 continues, it could drive substantial performance gains, possibly even if the size, training compute and training data of models do not increase.

However, it is uncertain the extent to which this pace of algorithmic progress can persist. If the most impactful and readily identifiable algorithmic advancements have already been implemented, algorithmic progress could slow. Alternatively, new innovations as impactful as the transformer architecture could be identified and allow continued or even more rapid capabilities growth. The rise of reasoning models trained using reinforcement learning points to continued opportunities for algorithmic advancements to drive substantial capability gains (Ho and Berg, 2025[152]).

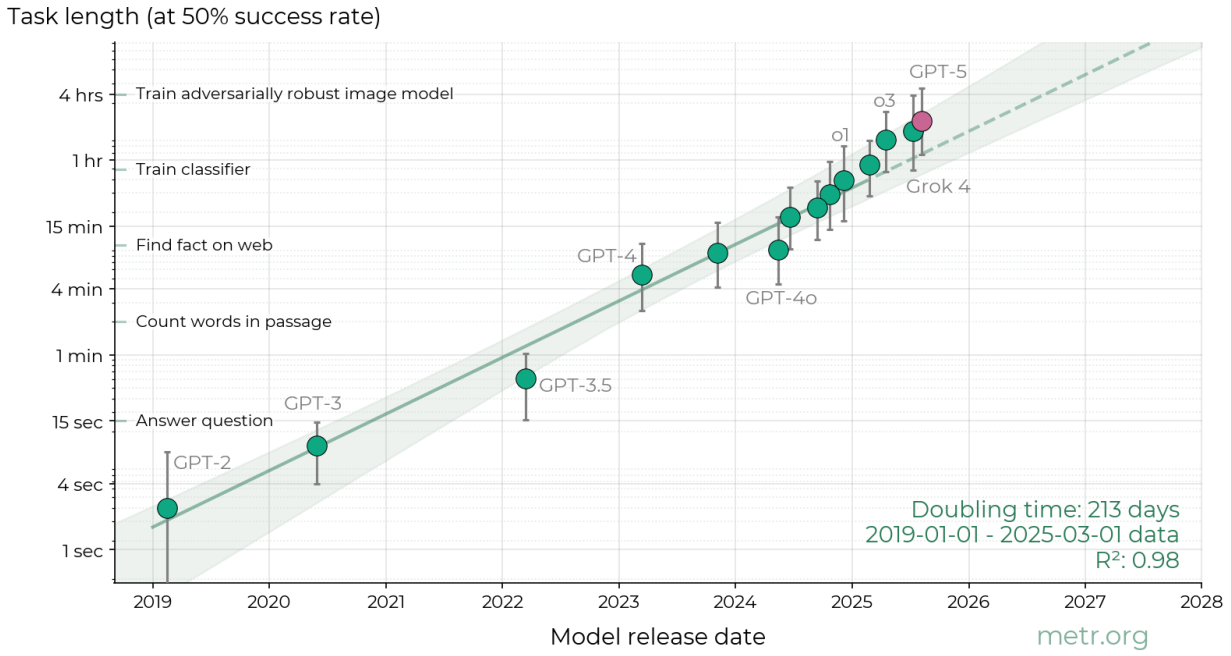
It is also uncertain the extent to which rapid algorithmic advancement is dependent upon continued compute scaling. Some analysis suggests that the most transformative of recent algorithmic innovations yield minimal or no benefit in small-scale AI systems, but unlock large improvements when models are scaled up (Josephson, 2025[50]). This suggests that future algorithmic progress may be partially dependent upon an ability to continue scaling compute. However, other analysis suggests that limits to compute are unlikely to dramatically slow AI algorithmic progress (Barnett, 2025[51]).

Annex D. Trend Extrapolations

D.1. The described scenarios approximately map onto different extrapolations of an observed exponential trend in the length of tasks that AI systems can complete with a 50% success rate

Some research has suggested that the length of tasks (as measured by the time it takes a human to complete the task) that AI systems can complete is doubling at a consistent rate (Kwa et al., 2025[130]). This research focuses on task types for which task length has a robust correlation with task difficulty, allowing task length to be used as a proxy for task difficulty. In software engineering, for which there are the best available data, the length of tasks that AI can perform appear to be doubling approximately every seven months (Figure 5).

Figure 5. Length of software engineering tasks that AI systems can autonomously complete with a 50% success rate has doubled every seven months



Note: This work is made available under the Creative Commons Attribution 4.0 International licence.

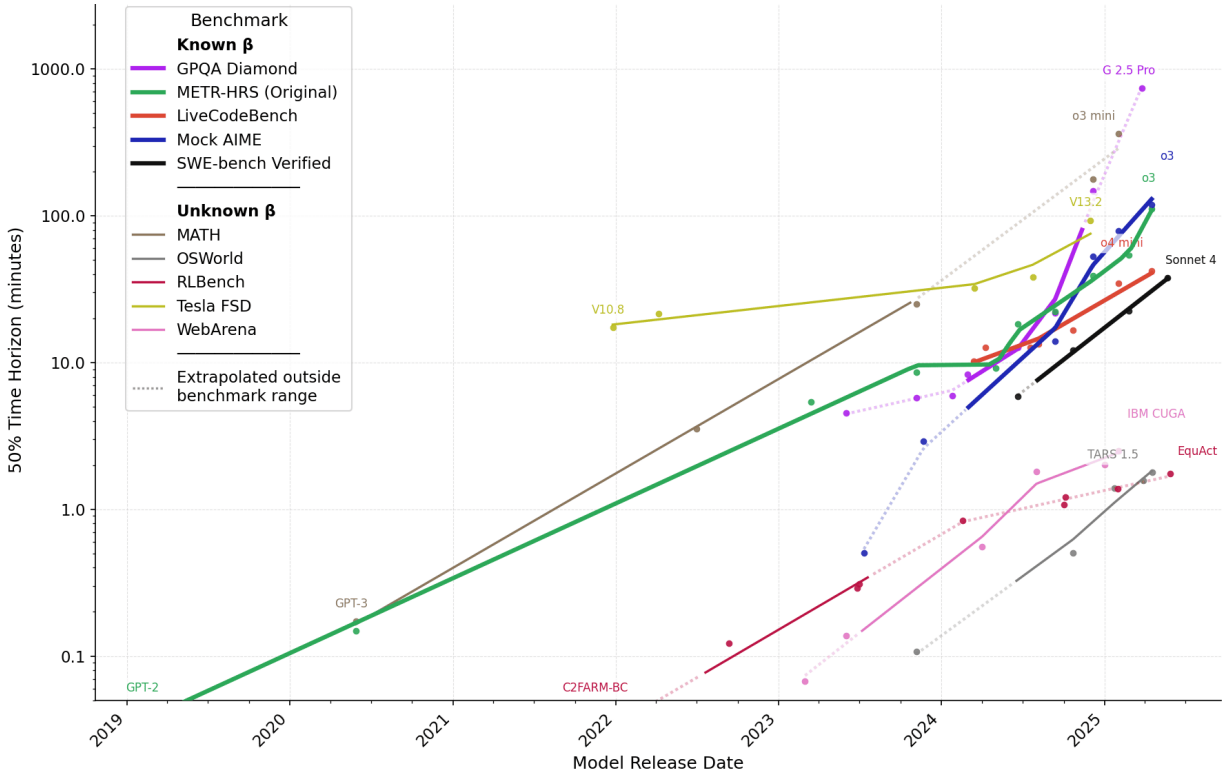
Source: (METR, 2025[155])

The length of software engineering tasks that AI systems can complete with an 80% success rate is shorter, with GPT-5 able to complete 217-minute-long software engineering tasks at a 50% success rate, and 26-

minute tasks with an 80% success rate. However, the ability to complete long software engineering tasks with an 80% success rate is also increasing rapidly, with the same trend and doubling time of seven months.

Similar or faster doubling times have been observed in the length of tasks AI systems can complete in scientific question answering (QPQA Diamond), mathematics questions (Mock AIME), vision-guided object manipulation (RL Bench), and vision-guided computer use (OSWorld and WebArena), though this is based on fewer data (Figure 6). For real-world autonomous driving tasks (Tesla FSD), the length of tasks AI systems can reliably complete is increasing substantially more slowly, doubling every 1.7 years.

Figure 6. The length of tasks that AI systems can autonomously complete with a 50% success rate has been increasing for a range of task types



Note: This work is made available under the Creative Commons Attribution 4.0 International licence.

Source: (METR, 2025[8])

As one input to help inform these scenarios, each scenario includes a different extrapolation of these hypothesised exponential trends, based on the currently observed capability level and doubling time from a range of benchmarks transposed to a common metric of task length (Figure 7).

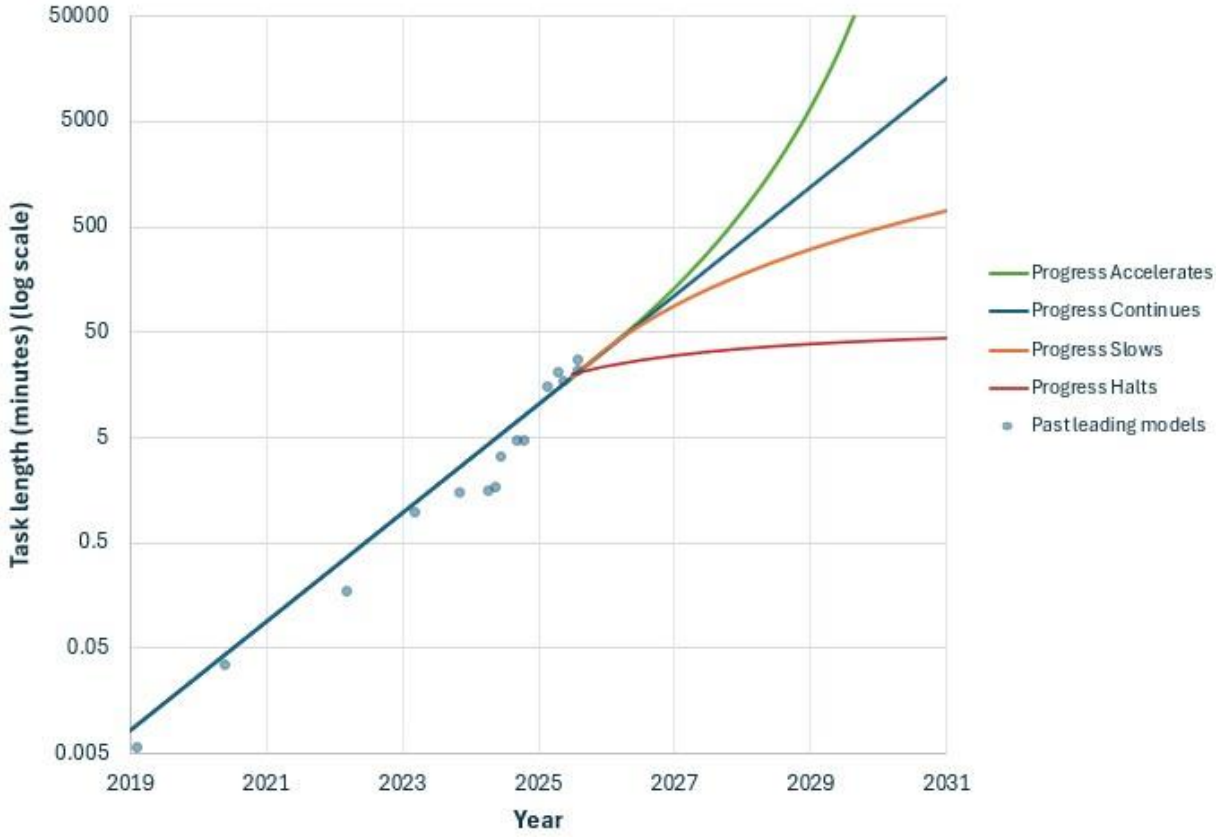
Table 14 Observed performance and rate of progress of AI systems on benchmarks assessing the length of tasks AI systems can complete with a 50% success rate across different domains

Task (benchmark)	Observed maximum task length with 50% success rate (minutes)	Observed doubling time in task length with 50% success rate
Scientific reasoning (GPQA Diamond)	741	2.5
Mathematical reasoning (Mock AIME)	119	2.9
Software engineering (METR-HRS)	137	7
Autonomous driving (Tesla FSD)	93	20
Computer use (OS World)	1.8	4
Web navigation (WebArena)	2.5	5
Simulated robotics (RLBench)	1.8	6

Source: (METR, 2025[8])

Figure 7 visualises the extrapolation of this trend for software engineering tasks, one of the seven task categories for which this extrapolation was conducted. The Progress Continues scenario assumes that the maximum task length that AI systems can complete with a 50% and 80% success rate continues doubling at currently observed rates. The Progress Halts scenario assumes AI capabilities will plateau at double the currently observed task length (which would be attained in six months if current trends continued). The Progress Slows scenario assumes that the doubling time will slow by 30% for each doubling, resulting in substantially sub-exponential growth in AI capabilities. The Progress Accelerates scenario assumes that the doubling time will speed up by 10% for each doubling, resulting in super-exponential growth in AI capabilities.

Figure 7. Four scenarios for future AI capabilities, visualised here by the length of software engineering tasks that leading AI systems can autonomously complete with an 80% success rate



Source: Authors' elaboration based on (Kwa et al., 2025[130]).

The exact projections in these scenarios are illustrative and intended to highlight the broad categories of different plausible outcomes for AI progress. For instance, in the Progress Accelerates scenario, the 10% acceleration figure is relatively arbitrary. If AI progress accelerates via more rapid breakthroughs, continued scaling, or AI-assisted coding, it is uncertain to what extent they will do so. In this projection, if AI software engineering assistants, continued scaling, or new algorithmic breakthroughs instead caused each doubling to occur 30% faster, then this scenario would come about in 2027. If each doubling only came 3% faster, it could come about in 2034. Even in a purely exponential scenario, without any acceleration of doubling times, this scenario could come about by 2037. However, if scaling continues at its current rate, projections suggest that limitations such as energy access could lessen the probability of AI continuing to advance at the same rapid pace beyond 2030 (Sevilla et al., 2024[49]).

This trend extrapolation was conducted for all seven task categories to help inform the development of these scenarios, with the results listed in the scenario descriptions as “Indicative capabilities in mid-2030”.[5] While these task categories do not directly correspond to the capability categories used in the OECD AI Capability Indicators, these indicative capability levels provide an additional way of conceptualising plausible AI capabilities in 2030.

This analysis does not extrapolate to task lengths beyond one month. It assumes that beyond one month task lengths, the time horizon metric breaks down as a useful indicator of AI capabilities. To represent this,

Figure 7 does not extrapolate beyond a task length of 43,830 minutes (roughly one month) and any task length projected to be longer than one month in the scenarios is described as “>1 month”.

These indicative trend extrapolations were one input among many used to validate and inform the scenarios. The key inputs to the construction of the scenarios were the capability trends and uncertainties identified in Chapters 2 and 3, a review of the relevant literature, and interviews and validation with leading AI experts.

References

Alam, M. and N. Rastogi (2025), “Limits of Generalization in RLVR: Two Case Studies in Mathematical Reasoning”, *arXiv*, https://doi.org/10.48550/arXiv.2510.27044. [37]

Al-Maamari, A. (2025), “Can You Trust Your Copilot? A Privacy Scorecard for AI Coding Assistants”, *arXiv*, https://doi.org/10.48550/arXiv.2509.20388. [57]

Altman, S. et al. (2025), *Pre-Training GPT-4.5*, OpenAI, https://www.youtube.com/watch?v=6nJZopACRuQ. [36]

Anthropic (2025), *How Anthropic teams use Claude Code*, https://www.anthropic.com/news/how-anthropic-teams-use-claude-code. [59]

Associated Press (2024), *Google says it will rethink its plans for a big data center in Chile over water worries*, Associated Press, https://apnews.com/article/chile-google-data-center-water-drought-environment-d1c6a7a8e8e6e45257ac84fb750b2162. [66]

Bahri, Y. et al. (2024), “Explaining Neural Scaling Laws”, *arXiv*, https://doi.org/10.1073/pnas.2311878121. [19]

Bai, A. et al. (2022), “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback”, *arXiv*, https://doi.org/10.48550/arXiv.2204.05862. [147]

Barnett, P. (2025), *Compute Requirements for Algorithmic Innovation in Frontier AI Models*, https://doi.org/10.48550/arXiv.2507.10618. [51]

Becker, J. et al. (2025), “Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity”, *arXiv*, https://doi.org/10.48550/arXiv.2507.09089. [62]

Bloom, N. et al. (2020), “Are Ideas Getting Harder to Find?”, *American Economic Review*, Vol. 110/4, pp. 1104–44, https://doi.org/10.1257/aer.20180338. [77]

Borg, M. et al. (2025), “Echoes of AI: Investigating the Downstream Effects of AI Assistance on Software Maintainability”, *arXiv*, https://doi.org/10.48550/arXiv.2507.00788. [61]

Bort, J. (2025), *OpenAI priced GPT-5 so low, it may spark a price war*, https://techcrunch.com/2025/08/08/openai-priced-gpt-5-so-low-it-may-spark-a-price-war/. [133]

Brill, A. (2024), “Neural Scaling Laws Rooted in the Data Distribution”, *arXiv*, https://doi.org/10.48550/arXiv.2412.07942. [20]

Caballero, E. et al. (2023), “Broken Neural Scaling Laws”, *arXiv*, https://doi.org/10.48550/arXiv.2210.14891. [23]

Chen, M. et al. (2021), “Evaluating Large Language Models Trained on Code”, *arXiv*, https://doi.org/10.48550/arXiv.2107.03374. [91]

Chen, Z. et al. (2025), “Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies”, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23881–23899, https://doi.org/10.18653/v1/2025.acl-long.1163. [22]

Chiang, W. et al. (2024), *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, https://doi.org/10.48550/arXiv.2403.04132. [120]

Chi, H. et al. (2024), *Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?*, https://doi.org/10.48550/arXiv.2506.21215. [101]

Child, R. et al. (2019), “Generating Long Sequences with Sparse Transformers”, *arXiv*, https://doi.org/10.48550/arXiv.1904.10509. [149]

Chollet, F. et al. (2025), “ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems”, *arXiv*, https://doi.org/10.48550/arXiv.2505.11831. [100]

Christiano, P. et al. (2017), “Deep reinforcement learning from human preferences”, *arXiv*, https://doi.org/10.48550/arXiv.1706.03741. [146]

Cobbe, K. et al. (2021), “Training Verifiers to Solve Math Word Problems”, *arXiv*, https://doi.org/10.48550/arXiv.2110.14168. [90]

Cotroneo, D., I. Cristina and P. Liguori (2025), “Human-Written vs. AI-Generated Code: A Large-Scale Study of Defects, Vulnerabilities, and Complexity”, *arXiv*, https://doi.org/10.48550/arXiv.2508.21634. [56]

Cui, Z. et al. (2025), “The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers”, *SSRN*, https://doi.org/10.2139/ssrn.4945566. [60]

D’Ambrosio, D. et al. (2024), “Achieving Human Level Competitive Robot Table Tennis”, *arXiv*, https://doi.org/10.48550/arXiv.2408.0390. [109]

Daniotti, S. et al. (2026), “Who is using AI to code? Global diffusion and impact of generative AI”, *Science*, [preprint version accessed in October 2025], https://doi.org/10.1126/science.adz9311. [54]

de Rooij, A. and M. Biskjaer (2025), “Has AI Surpassed Humans in Creative Idea Generation? A Meta-Analysis”, *36th Annual Conference of the European Association of Cognitive Ergonomics*, https://doi.org/10.31234/osf.io/9u2ke_v2. [107]

DeepSeek (2025), “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning”, https://doi.org/10.48550/arXiv.2501.12948. [35]

Desdevises, J. (2025), “The paradox of creativity in generative AI: high performance, human-like bias, and limited differential evaluation”, *Frontiers in Psychology*, Vol. 16, https://doi.org/10.3389/fpsyg.2025.1628486. [106]

Deshmukh, N. et al. (1998), *Resegmentation of Switchboard*, https://www.isca-archive.org/icslp_1998/deshmukh98_icslp.pdf. [86]

Dohmatob, E. et al. (2024), "A Tale of Tails: Model Collapse as a Change of Scaling Laws", *arXiv*, https://doi.org/10.48550/arXiv.2402.07043. [21]

Du, C. et al. (2025), "Human-like object concept representations emerge naturally in multimodal large language models", *Nature Machine Intelligence*, Vol. 7, pp. 860-875, https://doi.org/10.1038/s42256-025-01049-z. [27]

Elyoseph, Z. et al. (2024), "Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study", *JMIR Mental Health*, p. e54369, https://doi.org/10.2196/54369. [111]

Emberson, L. and J. You (2025), *GPT-5 and GPT-4 were both major leaps in benchmarks from the previous generation*, https://epoch.ai/data-insights/gpt-capabilities-progress. [131]

Epoch AI (2025), *AI Benchmarking Hub*, https://epoch.ai/benchmarks (accessed on 2025). [4]

Epoch AI (2025), *Data on Notable AI Models*, https://epoch.ai/data/ai-models (accessed on 2025). [18]

Eriksson, M. et al. (2025), "Can we trust AI benchmarks? An interdisciplinary review of current issues in AI evaluation", *arXiv*, https://doi.org/10.48550/arXiv.2502.06559. [117]

Fodor, J. (2025), "Line Goes Up? Inherent Limitations of Benchmarks for Evaluating Large Language Models", *arXiv*, https://doi.org/10.48550/arXiv.2502.14318. [118]

GitHub (2025), *AI in Softeware Development 2024 Survey*, GitHub, https://github.blog/news-insights/research/survey-ai-wave-grows/. [53]

Glazer, E. et al. (2024), "FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI", *arXiv*, https://doi.org/10.48550/arXiv.2411.04872. [94]

Goldman Sachs (2025), *How AI is Transforming Data Centers and Ramping Up Power Demand*, https://www.goldmansachs.com/insights/articles/how-ai-is-transforming-data-centers-and-ramping-up-power-demand. [73]

Google (2025), *Long context*, https://ai.google.dev/gemini-api/docs/long-context. [99]

Google Deepmind (2025), *AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms*, https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/. [47]

Google Deepmind (2025), *Gemini: A Family of Highly Capable Multimodal Models*, https://doi.org/10.48550/arXiv.2312.11805. [138]

Gujju, Y. and A. Matsuo (2024), "Quantum Machine Learning on Near-Term Quantum Devices: Current State of Supervised and Unsupervised Techniques for Real-World Applications", *arXiv*, https://doi.org/10.48550/arXiv.2307.00908. [140]

Hendrycks, D. et al. (2020), "Measuring Massive Multitask Language Understanding", *arXiv*, https://doi.org/10.48550/arXiv.2009.03300. [89]

Hernandez, D. and T. Brown (2020), "Measuring the Algorithmic Efficiency of Neural Networks", *arXiv*, https://doi.org/10.48550/arXiv.2005.04305. [154]

Hestness, J. et al. (2017), "Deep Learning Scaling is Predictable, Empirically", *arXiv*, https://doi.org/10.48550/arXiv.1712.00409. [13]

Heyman, A. and J. Zylberberg (2025), "Evaluating the Systematic Reasoning Abilities of Large Language Models through Graph Coloring", *arXiv*, https://doi.org/10.48550/arXiv.2502.07087. [28]

Ho, A. and A. Berg (2025), "Quantifying the algorithmic improvement from reasoning models", *Gradient Updates*, https://epoch.ai/gradient-updates/quantifying-the-algorithmic-improvement-from-reasoning-models. [152]

Ho, A. et al. (2024), "Algorithmic progress in language models", *arXiv*, https://doi.org/10.48550/arXiv.2403.05812. [48]

Hou, W. and Z. Ji (2024), "Comparing Large Language Models and Human Programmers for Generating Programming Code", *Advanced Science*, Vol. 12/8, p. 2412279, https://doi.org/10.1002/advs.202412279. [3]

Huang, S. et al. (2024), "AI Technology panic—is AI Dependence Bad for Mental Health? A Cross-Lagged Panel Model and the Mediating Roles of Motivations for AI Use Among Adolescents", *Psychology Research and Behavior Management*, Vol. 17, pp. 1087-1102, https://doi.org/10.2147/PRBM.S440889. [116]

Hubert, K., K. Awa and D. Zabelina (2025), *Generative artificial intelligence models outperform students on divergent and convergent thinking assessments*, Research Square, https://doi.org/10.21203/rs.3.rs-6762129/v1. [102]

International Energy Agency (2025), *Energy and AI*, https://www.iea.org/reports/energy-and-ai. [69]

IPCC (1999), *Aviation and the Global Atmosphere*, Cambridge University Press, https://archive.ipcc.ch/ipccreports/sres/aviation/index.php?idp=92. [70]

Jayatunga, M. et al. (2024), "How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons", *Drug Discovery Today*, Vol. 29/6, https://doi.org/10.1016/j.drudis.2024.104009. [97]

Jones, C. and B. Bergen (2025), "Large Language Models Pass the Turing Test", *arXiv*, https://doi.org/10.48550/arXiv.2503.23674. [110]

Josephson, H. (2025), *How fast can algorithms advance capabilities?*, https://epoch.ai/gradient-updates/how-fast-can-algorithms-advance-capabilities. [50]

Kahn, J. (2025), "The $19.6 billion pivot: How OpenAI's 2-year struggle to launch GPT-5 revealed that its core AI strategy has stopped working", *Fortune*, https://fortune.com/2025/02/25/what-happened-gpt-5-openai-orion-pivot-scaling-pre-training-llm-agi-reasoning/. [125]

Kaplan, J. et al. (2020), *Scaling Laws for Neural Language Models*, OpenAI, https://doi.org/10.48550/arXiv.2001.08361. [14]

Kazemi, M. et al. (2025), "BIG-Bench Extra Hard", *arXiv*, https://doi.org/10.48550/arXiv.2502.19187. [5]

Khalid, I., A. Nourollah and S. Schockaert (2025), "Benchmarking Systematic Relational Reasoning with Large Language and Reasoning Models", *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8843–8869, https://doi.org/10.18653/v1/2025.acl-long.433. [29]

Kiela, D. et al. (2023), *Plotting Progress in AI (Update of data from Kiela, D. et al. (2021) Dynabench: Rethinking Benchmarking in NLP, arXiv, https://doi.org/10.48550/arXiv.2104.14337)*, https://contextual.ai/blog/plotting-progress. [80]

Klienman, Z. and B. Wheeler (2025), *Concern UK's AI ambitions could lead to water shortages*, https://www.bbc.co.uk/news/articles/ce85wx9jjndo. [67]

Krizhevsky, A., I. Sutskever and G. Hinton (2012), "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25*, Vol. 60/6, pp. 84-90, https://doi.org/10.1145/306538. [123]

Kwa, T. et al. (2025), "Measuring AI Ability to Complete Long Tasks", *arXiv*, https://doi.org/10.48550/arXiv.2503.14499. [130]

LeCun, Y. and C. Cortes (2010), "The MNIST Database of Handwritten Digits", https://www.lri.fr/~marc/Master2/MNIST_doc.pdf. [85]

Lee, D. et al. (2025), "Bayesian Neural Scaling Law Extrapolation with Prior-Data Fitted Networks", *arXiv*, https://doi.org/10.48550/arXiv.2505.23032. [15]

Leeman, J. et al. (2024), "Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis", *PRX Energy*, Vol. 3/1, p. 011002, https://doi.org/10.1103/PRXEnergy.3.011002. [98]

Lewis, K. (2020), "The Science of Antibiotic Discover", *Cell*, Vol. 181/1, https://doi.org/10.1016/j.cell.2020.02.056. [71]

Li, G. et al. (2025), *The Developments and Challenges towards Dexterous and Embodied Robotic Manipulation: A Survey*, https://doi.org/10.48550/arXiv.2507.11840. [39]

Lin, H. and H. Cheng (2025), *Gemini achieves gold-medal level at the International Collegiate Programming Contest World Finals*, https://deepmind.google/discover/blog/gemini-achieves-gold-level-performance-at-the-international-collegiate-programming-contest-world-finals/. [7]

Liu, K. et al. (2025), "Reinforcement Learning Meets Large Language Models: A Survey of Advancements and Applications Across the LLM Lifecycle", *arXiv*, https://doi.org/10.48550/arXiv.2509.16679. [34]

Liu, Y. et al. (2024), "Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI", *arXiv*, https://doi.org/10.48550/arXiv.2407.06886. [40]

LMArena (2025), *LMArena Text Arena*, https://lmarena.ai/leaderboard/text (accessed on 24 September 2025). [121]

Maslej, N. et al. (2025), *Artificial Intelligence Index Report 2025*, Institute for Human-Centered AI, https://doi.org/10.48550/arXiv.2504.07139. [153]

Methur, L. et al. (2025), "Social Genome: Grounded Social Reasoning Abilities of Multimodal Models", *arXiv*, https://doi.org/10.48550/arXiv.2502.15109. [113]

METR (2025), *Details about METR's evaluation of OpenAI GPT-5*, https://metr.github.io/autonomy-evals-guide/gpt-5-report/. [155]

METR (2025), *How Does Time Horizon Vary Across Domains?*, https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/. [8]

Metz, C. (2025), "Google A.I. System Wins Gold Medal in International Math Olympiad", *The New York Times*, https://www.nytimes.com/2025/07/21/technology/google-ai-international-mathematics-olympiad.html (accessed on 2025). [6]

Metz, C. (2016), *Google's AI Wins Pivotal Second Game in Match With Go Grandmaster*, https://www.wired.com/2016/03/googles-ai-wins-pivotal-game-two-match-go-grandmaster/. [46]

Nawrot, P. et al. (2025), "The Sparse Frontier: Sparse Attention Trade-offs in Transformer LLMs", *arXiv*, https://doi.org/10.48550/arXiv.2504.17768. [150]

Ni, R. et al. (2025), "Benchmarking and understanding compositional relational reasoning of LLMs", *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, https://doi.org/10.1609/aaai.v39i18.34170. [24]

Novikov, A. et al. (2025), *AlphaEvolve: A coding agent for scientific and algorithmic discovery*, https://doi.org/10.48550/arXiv.2506.13131. [108]

OECD (2025), *A quantum technologies policy primer*, OECD Publishing, https://doi.org/10.1787/fd1153c3-en. [65]

OECD (2025), *Catalogue of Tools & Metrics for Trustworthy AI*, https://oecd.ai/en/catalogue/metrics. [122]

OECD (2025), *Introducing the OECD AI Capability Indicators*, OECD Publishing, https://doi.org/10.1787/be745f04-en. [1]

OECD (2024), *Futures of Global AI Governance; Co-Creating an Approach for Transforming Economies and Societies*, OECD Publishing, https://www.oecd.org/content/dam/oecd/en/about/programmes/strategic-foresight/GSG%20Background%20Note_GSG(2024)1en.pdf. [63]

OpenAI (2025), *API Pricing (21 March 2025)*, http://web.archive.org/web/20250321002735/https://openai.com/api/pricing/. [136]

OpenAI (2025), *API Pricing (24 September 2025)*, https://openai.com/api/pricing/. [134]

OpenAI (2025), *API Pricing (28 March 2025)*, http://web.archive.org/web/20250328101754/https://openai.com/api/pricing/. [127]

OpenAI (2025), *ChatGPT — Release Notes*, https://help.openai.com/en/articles/6825453-chatgpt-release-notes. [132]

OpenAI (2025), "Competitive Programming with Large Reasoning Models", *arXiv*, https://doi.org/10.48550/arXiv.2502.06807. [33]

OpenAI (2025), *Introducing ChatGPT agent: bridging research and action*, https://openai.com/index/introducing-chatgpt-agent/. [44]

OpenAI (2025), *Introducing GPT-4.5*, https://openai.com/index/introducing-gpt-4-5/. [128]

OpenAI (2025), *Introducing GPT-5*, https://openai.com/index/introducing-gpt-5/. [129]

OpenAI (2025), *July 23 2025 - API Pricing*, http://web.archive.org/web/20250000000000*/https://openai.com/api/pricing/. [135]

OpenAI (2025), *Our Structure*, https://openai.com/our-structure/. [78]

OpenAI (2024), "GPT-4o System Card", *arXiv*, https://doi.org/10.48550/arXiv.2410.21276. [151]

OpenAI (2024), *Learning to reason with LLMs*, https://openai.com/index/learning-to-reason-with-llms/ (accessed on 2025). [30]

OpenAI (2023), *Planning for AGI and beyond*, https://openai.com/index/planning-for-agi-and-beyond/. [58]

Otten, D. et al. (2025), "Prompting in Practice: Investigating Software Developers' Use of Generative AI Tools", *arXiv*, https://doi.org/10.48550/arXiv.2510.06000. [55]

Ott, S. et al. (2022), "Mapping global dynamics of benchmark creation and saturation in artificial intelligence", *Nature Communications*, Vol. 13, p. 6793, https://doi.org/10.1038/s41467-022-34591-0. [79]

Ouyang, L. et al. (2022), "Training language models to follow instructions with human feedback", *arXiv*, https://doi.org/10.48550/arXiv.2203.02155. [144]

Owen, D. (2024), "How Predictable Is Language Model Benchmark Performance?", *arXiv*, https://doi.org/10.48550/arXiv.2401.04757. [17]

Paglieri, D., B. Cupial and U. Piterbarg (2024), "BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games", *arXiv*, https://doi.org/10.48550/arXiv.2411.13543. [16]

Palazzolo, S., E. Woo and A. Efrati (2025), *Inside OpenAI's Rocky Path to GPT-5*, https://www.theinformation.com/articles/inside-openais-rocky-path-gpt-5. [126]

Patwardhan, T. et al. (2025), *GDPval: Evaluating AI model performance on real-world economically valuable tasks*, OpenAI, https://cdn.openai.com/pdf/d5eb7428-c4e9-4a33-bd86-86dd4bcf12ce/GDPval.pdf. [10]

Prabhakar, A., T. Griffiths and T. McCoy (2024), "Deciphering the Factors Influencing the Efficacy of Chain-of-Thought: Probability, Memorization, and Noisy Reasoning", *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3710–3724, https://doi.org/10.48550/arXiv.2407.01687. [25]

Proietti, L., S. Perrelle and R. Navigli (2025), "Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress", *arXiv*, https://doi.org/10.48550/arXiv.2506.19571. [9]

Rahman, R. (2024), *Performance per dollar improves around 30% each year*, EpochAI, https://epoch.ai/data-insights/price-performance-hardware. [137]

Rajpurkar, P. et al. (2016), "SQuAD: 100,000+ Questions for Machine Comprehension of Text", *arXiv*, https://doi.org/10.48550/arXiv.1606.05250. [81]

Refoua, E. et al. (2025), "The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli", *Computers in Human Behavior Reports*, Vol. 19, https://doi.org/10.1016/j.chbr.2025.100702. [112]

Robinson, D. and D. Nadal (2025), *Synthetic biology in focus: Policy issues and opportunities in engineering life*, OECD Publishing, https://doi.org/10.1787/3e6510cf-en. [64]

Roser, M., H. Ritchie and E. Mathieu (2023), *What is Moore's Law?*, https://archive.ourworldindata.org/20251230-000121/moores-law.html. [74]

Russakovsky, O. et al. (2015), "ImageNet Large Scale Visual Recognition Challenge", *IJCV*, Vol. 115, pp. 211–252, https://doi.org/10.1007/s11263-015-0816-y. [82]

Sarkar, S. and R. Alqasemi (2025), "Neural Interfaces for Robotics and Prosthetics: Current Trends", *Journal of Sensor and Actuator Networks*, Vol. 14/5, p. 105, https://doi.org/10.3390/jsan14060105. [41]

Saxon, M. et al. (2024), "Benchmarks as Microscopes: A Call for Model Metrology", *arXiv*, https://doi.org/10.48550/arXiv.2407.16711. [119]

Sevilla, J. et al. (2024), *Can AI Scaling Continue Through 2030?*, https://epoch.ai/blog/can-ai-scaling-continue-through-2030. [49]

Sevilla, J. et al. (2022), "Compute Trends Across Three Eras of Machine Learning", *arXiv*, https://doi.org/10.1109/IJCNN55064.2022.9891914. [124]

Shaw, K., A. Agarwal and D. Pathak (2023), "LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand", *arXiv*, https://doi.org/10.48550/arXiv.2309.06440. [43]

Shazeer, N. et al. (2017), "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", *arXiv*, https://doi.org/10.48550/arXiv.1701.06538. [148]

Shumailov, I. et al. (2024), "AI models collapse when trained on recursively generated data", *Nature*, Vol. 631, pp. 755–759, https://doi.org/10.1038/s41586-024-07566-y. [72]

Solomon, S. (2024), *Sustainable by design: Next-generation datacenters consume zero water for cooling*, https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/12/09/sustainable-by-design-next-generation-datacenters-consume-zero-water-for-cooling/. [68]

Song, P., P. Han and N. Goodman (2025), "A Survey on Large Language Model Reasoning Failures", *2nd AI for Math Workshop @ ICML 2025*, https://openreview.net/forum?id=hsgMn4KBFG. [11]

Srivastava, A. et al. (2022), "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models", *arXiv*, https://doi.org/10.48550/arXiv.2206.04615. [93]

Stackoverflow (2025), *2025 Developer Survey*, https://survey.stackoverflow.co/2025/. [52]

Sun, L. et al. (2025), "Large language models show both individual and collective creativity comparable to humans", *Thinking Skills and Creativity*, Vol. 57, p. 101870, https://doi.org/10.1016/j.tsc.2025.101870. [103]

Su, Y. et al. (2025), “Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains”, *arXiv*, https://doi.org/10.48550/arXiv.2503.23829. [76]

Suzgun, M. et al. (2022), “Challenging BIG-Bench tasks and”, *arXiv*, https://doi.org/10.48550/arXiv.2210.09261. [92]

The Royal Swedish Academy of Sciences (2024), *They cracked the code for proteins' amazing structures*, https://www.nobelprize.org/prizes/chemistry/2024/press-release/. [96]

Tran, K. et al. (2025), “Multi-Agent Collaboration Mechanisms: A Survey of LLMs”, *arXiv*, https://doi.org/10.48550/arXiv.2501.06322. [45]

Ünlütabak, B. and O. Bal (2025), “Theory of mind performance of large language models: A comparative analysis of Turkish and English”, *Computer Speech & Language*, Vol. 89, https://doi.org/10.1016/j.csl.2024.101698. [114]

Urrea, C. and J. Kern (2025), “Recent Advances and Challenges in Industrial Robotics: A Systematic Review of Technological Trends and Emerging Applications”, *Processes*, Vol. 13, p. 832, https://doi.org/10.3390/pr13030832. [42]

Vaswani, A. et al. (2017), “Attention Is All You Need”, *31st Conference on Neural Information Processing Systems*, pp. 6000–6010. [143]

Wang, A. et al. (2019), “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, *arXiv*, https://doi.org/10.48550/arXiv.1905.00537. [84]

Wang, A. et al. (2018), “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”, *arXiv*, https://doi.org/10.48550/arXiv.1804.07461. [83]

Wang, J. et al. (2024), “Evaluating and Modeling Social Intelligence: A Comparative Study of Human and AI Capabilities”, *arXiv*, https://doi.org/10.48550/arXiv.2405.11841. [115]

Wei, J. et al. (2023), “Chain-of-Thought Prompting Elicits Reasoning”, *arXiv*, https://doi.org/10.48550/arXiv.2201.11903. [31]

Wetterstrand, K. (2022), *DNA Sequencing Costs: Data*, National Human Genome Research Institute, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. [75]

Xiong, W. et al. (2017), *Achieving Human Parity in Conversational Speech Recognition*, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/ms_parity.pdf. [87]

Xuan, W. et al. (2025), “MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation”, *arXiv*, https://doi.org/10.48550/arXiv.2503.10497. [12]

Yang, Y. et al. (2025), “Emergent Symbolic Mechanisms Support Abstract Reasoning in Large Language Models”, *arXiv*, https://doi.org/10.48550/arXiv.2502.20332. [26]

Yeadon, W. et al. (2025), “Evaluating AI and human authorship quality in academic writing through physics essays”, *European Journal of Physics*, Vol. 45/5, p. 055703, https://doi.org/10.1088/1361-6404/ad669d. [2]

Yin, Z. et al. (2025), “Experimental quantum-enhanced kernel-based machine learning on a photonic processor”, *Nature Photonics*, Vol. 19, pp. 1020–1027, https://doi.org/10.1038/s41566-025-01682-5. [142]

You, J. (2025), *How far can reasoning models scale?*, https://epoch.ai/gradient-updates/how-far-can-reasoning-models-scale. [139]

Yue, Y. et al. (2025), "Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?", *arXiv*, https://doi.org/10.48550/arXiv.2504.13837. [32]

Zellers, R. et al. (2019), "HellaSwag: Can a Machine Really Finish Your Sentence?", *arXiv*, https://doi.org/10.48550/arXiv.1905.07830. [88]

Zhang, M. et al. (2025), "AI Delivers Creative Output but Struggles with Thinking Processes", *arXiv*, https://doi.org/10.48550/arXiv.2503.23327. [104]

Zhang, Y. et al. (2025), "NoveltyBench: Evaluating Language Models for Humanlike Diversity", *arXiv*, https://doi.org/10.48550/arXiv.2504.05228. [105]

Zhao, H. and D. Deng (2025), "Entanglement-induced provable and robust quantum learning advantages", *npj Quantum Information*, Vol. 11, p. 127, https://doi.org/10.1038/s41534-025-01078-x. [141]

Zhao, W., K. Gangaraju and F. Yuan (2025), "Multimodal perception-driven decision-making for human-robot interaction: a survey", *Frontiers in Robotics and AI*, Vol. 12, https://doi.org/10.3389/frobt.2025.1604472. [38]

Zheng, Z. et al. (2025), "LiveCodeBench Pro: How Do Olympiad Medalists Judge LLMs in Competitive Programming?", *arXiv*, https://doi.org/10.48550/arXiv.2506.11928. [95]

Ziegler, D. et al. (2019), *Fine-Tuning Language Models from Human Preferences*, https://doi.org/10.48550/arXiv.1909.08593. [145]

Endnotes

[1] The tasks included were self-contained, one shot, precisely-specified, and did not require extensive tacit knowledge, access to personally identifiable information, use of proprietary software tools, interactivity, or communication between individuals.

[2] Pretraining is an initial stage of training where an AI system is trained on large amounts of data to develop a range of generic capabilities, before any more specific training or fine-tuning.

[3] For the methodology related to trend extrapolation used as input to these scenarios, see Annex D.

[4] Note that this scenario explores AI systems with undesirable deficiencies in memory and learning, distinct from beneficial forms of “forgetting” that are studied in the field of machine unlearning. Machine unlearning seeks to help grant individuals greater control over their personal data and an ability to delete it from AI systems.

[5] This extrapolation for other task categories was only conducted for 50% success rate task horizons, due to data availability. If current relationships hold, task horizons with an 80% success rate would be five to ten times shorter than those for 50% success rate.